TOWARDS THE GENESIS OF NEURONAL REGULATORY CATALOGS AND THEIR VOCABULARIES

by
Xylena Reed

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, MD

January, 2015

Abstract

My research has focused on the identification and characterization of the transcriptional enhancers required for the development and specification of neuronal populations that are impacted in human health and disease. The enhancer field has evolved quickly and as a result I have used multiple methods to identify cell-type restricted enhancers. I started by using sequence conservation to identify sequences flanking LMX1A and LMX1B that drive expression in the central nervous system. We identified 47/71 constructs driving reporter expression in the CNS of mosaic zebrafish embryos. I then identified multiple founders driving consistent expression overlapping with endogenous expression patterns in 22/47 stable lines. This is a good method for locus specific analysis but is not easily scalable to genome-wide enhancer identification.

In order to identify enhancers in a broader context we applied machine learning to create a classifier that identifies hindbrain enhancers genome-wide. Using a set of experimentally proven enhancers that drive expression in the hindbrain as a training set for a machine learning algorithm that searches for over representation of known transcription factor binding sites and *de novo* motifs, we predicted 40,000 hindbrain enhancers. The *in vivo* validation rate for tested elements reached 88% for expression in the hindbrain, displaying high sensitivity but low specificity. We attribute the lack of specificity to the heterogeneity of the training set and determined to employ a new approach to acquire a more homogenous cell population.

Previous work has established that the joint analysis of transcriptional co-activator EP300 with histone modification H3K4me1 by ChIP-seq in cultured cells yields a highly accurate catalog of putative enhancers. However, my specific neuronal subtypes of interest

are not obtainable in the large numbers necessary for EP300 ChIP-seq. Instead, I examined public data from human *substantia nigra* and worked to optimize a small-cell number ChIP-seq protocol for the analysis of *ex vivo* sorted neurons. I have completed histone ChIP-seq in sorted neurons from a transgenic mouse line driving EGFP in DA neurons. This work identifies a catalog of putative enhancers that may play important roles in the expression of genes required for the development and maintenance of DA neurons.

Preface

I am very grateful to have had the opportunity to complete PhD in the Insitute of Genetics Medicine with Andy McCallion as my preceptor. It has been a long and difficult journey but I have learned far more than I ever thought possible. Andy has been an excellent mentor and has come to my rescue many times when I have found myself floundering. I owe Andy many thanks for all that he has taught me about science and life.

I would also like to thank everyone who has passed through the McCallion lab since I have been here. Dave Gorkin laid the groundwork for much of my ChIP-seq work, and trained me in many techniques. Samantha Maragh was always warm and caring, and will forever be the in situ queen of the McCallion lab. Kipper Fletez-Brant rescued me from various computer problems and always came equipped with a good sense of humor. Greg Brzynski started project that became a large part of my thesis. Maggie Baker and Courtney Woods have injected new life into the lab. Sarah McClymont I trust completely to complete any loose ends I have left and I expect to be hearing wonderful things about her science in the future. All lab members, it has been a wonderful experience working with you and getting to know you over the years.

I thank my thesis committee Roger Reeves, Mike Beer, and particularly Dimitri and Andy for being my thesis readers. Your advice and input throught my graduate career has been invaluable. I also want to thank Dr. Valle and Kirby for heading such a wonderful graduate program, and Sandy for keeping everything running smoothly. I really appreciate all the help (and candy) that you have given me. I also thank all of the faculty, fellow graduate students, and other IGM members that have made this an unforgettable experience.

I would like to thank all of my family and friends who have supported me over the last 6 years. My parents, Rico and Terri, my brother, Zeno, and sister, Adar, have always supported me in doing whatever makes me happy even if they don't really understand it. They have all given my tremendous love and support throughout my life that made this possible. I could not have done this without the support of many friends who have came and gone throughout the years. Particularly Clarissa, Jenn, Pete, Raghu, Paul, Ciaran, Andre, and Chris who have given me indispensible scientific and life advice throughout my time here, and all of my other friends who have made my life enjoyable and full. Thank you to everyone who has been a part of my life and helped me to become who I am today.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

INTRODUCTION

1.1 Enhancers and transcriptional regulation

In general, the DNA content within an organism of each cell is identical but this identical genetic content is also responsible for the vast diversity of cell types and expression patterns that we see in all eukaryotic life forms. This diversity is achieved by closely coordinated and dynamic control of gene expression that allows for responses to subtle environmental and developmental cues. The first step in this regulation occurs as DNA is transcribed to RNA. Transcriptional regulation requires the complex orchestration of interactions between numerous nucleotide and protein components. Transcription factors (TFs) and co-activators bind to DNA and other proteins in order to either induce or repress gene transcription. Many of the protein components involved in this regulation have been well studied and their functions are well understood (Buratowski et al., 1989; Thomas and Chiang, 2010). However, the regulatory instructions encrypted at the noncoding DNA sequence level have remained difficult to decipher.

The best characterized of the noncoding sequence encoded components is the promoter. The promoter is usually found within the 100 base pairs (bp) of sequence surrounding the transcriptional start site (TSS) and acts as the assembly point for the basal transcriptional machinery (Burke and Kadonaga, 1997; Juven-Gershon and Kadonaga, 2010). There is no universal sequence to identify promoters, instead they are made up of a combination of sequence elements including the TATA box, the initiator (Inr), TFIIB Recognition Element (BRE), Downstream core Promoter Element (DPE) and Motif Ten Element (MTE). The TATA box is located -30 bp from the TSS and acts to bind TBP (TATA

Binding Protein). It was the first identified promoter sequence (Goldberg, 1979) but it was later determined to be present in only 10-15% of mammalian promoters (Carnici et al., 2006; Kim et al., 2005; Cooper et al., 2006). The Inr is the located at the TSS and is the most commonly identified promoter element (Smale and Baltimore, 1989; Ohler et al., 2002; Fitzgerlad et al., 2006; Gershenzon et al., 2006). BRE motifs are in a subset of promoters with TATA boxes (10-30%; Lagrange et al., 1998), while DPE motifs are generally present only in promoters that do not contain TATA boxes (Burke and Kadonaga, 1996; Ohler et al., 2002). Additionally the MTE motif can work independently or synergistically with any of the other core promoter sequences (Ohler et al., 2002; Lim et al., 2004). Despite the lack of precise sequence consensus of core promoter elements, promoters can still be readily identified by their position just 5' of genes through cDNA sequencing and via their interaction with the transcriptional machinery using chromatin immunoprecipitation (ChIP; Trinklein et al., 2004; Cooper et al., 2006).

A less well understood class of noncoding regulatory sequences includes enhancers and repressors and is characterized by the positive and negative modulation of gene expression. Enhancers are *cis* noncoding sequences that act to up-regulate gene expression, while repressors down-regulate expression in a cell-type-restricted manner. These sequences generally act independent of orientation and can be located upstream, or downstream of within the introns of their target gene. Some extreme examples have been shown to function from more than a megabase (Mb; Lettice et al., 2002, 2003) away from their target. Enhancers contain collections of transcription factor binding sites (TFBSs) that are predicted to function by binding TFs which in turn recruit EP300, CBP and mediator proteins (Figure 1-1; Panne et al., 2007; Noonan and McCallion, 2010; He et al., 2011). This complex of TFs and co-factors can then interact with the basal transcriptional machinery at the promoter of a

target gene to initiate transcription (Ong and Corces, 2011; Chepelev et al., 2012). Since enhancers and repressors are not found in a specific location relative to their target gene and there is no overt sequence characteristic of them historically, as a class they have been very difficult to identify. Additional factors involved in transcriptional regulation include chromatin structure, histone modifications and proximity of *cis* regulatory features to a gene. The histone modifications H3K4me1 and H3K27ac have been associated with enhancer function for many years (Heintzman et al. 2007; Rada-Iglesias et al. 2007; Heintzman et al., 2009; Creyghton et al., 2010).

1.2 Phenotypic effects of variation in human enhancers

The instructions encoded within regulatory elements, including enhancers, play important roles in directing cell fate determination, and allowing cells to respond to their environment by changing levels of gene expression. Variation in these elements is predicted to contribute significantly to normal human variation and disease risk. Mutations affecting enhancers have been known for many years to result in Mendelian inherited developmental defects in humans. However, our very limited understanding of the sequence basis of enhancers makes it difficult to identify relevant enhancer sequences from primary genomic sequence alone. As a result, our lab has an ongoing interest in identifying enhancer catalogs in a variety of cell types and determining the sequence vocabularies that make them up.

One of the earliest examples of a Mendelian disorder resulting from disruption of an enhancer was found in a familial form of aniridia, a developmental defect characterized by hypoplasia of the iris first described in the late 1950s (Shaw et al., 1960). Aniridia had been found to be caused by loss of function mutations in the coding region of *PAX6,* a TF with a paired box DNA binding domain involved in nervous system and eye development. Families

with aniridia were soon identified that had no coding mutations but did have cytogenetic rearrangements near the *PAX6* locus (Fantes et al., 1995). Breakpoint mapping using Yeast Artificial Chromosomes (YACs) implicated a region 125 kb downstream from *PAX6* as an enhancer important for regulating expression of PAX6 in the eye. Around the same time another group identified a sequence downstream of the quail *Pax-6* that drove expression specifically in neuro-retinal cells (Plaza et al., 1995). It was later shown that a single point mutation in this region causes loss of reporter expression in the retina of zebrafish and mice (Bhatia et al., 2013).

As a result of our constantly improving understanding of noncoding DNA the literature is now full of examples of both Mendelian and complex human diseases arising from alterations in predicted enhancers. Other diseases included within this list are mutations affecting an enhancer of *SHH* located 1 Mb away leading to pre-axial polydactyly (Lettice et al., 2002, 2003); *SOX9* enhancer mutations linked to Pierre Robin Sequence (Benko et al., 2009); *POU3F4* regulatory mutations associating with X-linked deafness type 3 (De Kok et al., 1995; Ahn et al., 2009; Naranjo et al., 2010); and sex-dependent mutations of an enhancer for *RET* causing forms of Hirschsprung disease (Emison et al., 2005). We hypothesize that as the tools to study enhancers continue to evolve we will identify many more phenotypic effects that are rooted in non-coding DNA. Evidence for this hypothesis is found in the hundreds of Genome-Wide Association Study (GWAS) results that implicate noncoding regions with a frequency of almost 95% (Maurano et al., 2012; Welter et al., 2014). We propose to leverage these GWAS results to identify functional putative enhancers with roles in health and disease.

1.3 Sequenced based analysis for the identification of enhancers

In the years following the sequencing of the human and mouse genomes the most commonly used method of enhancer identification became the use of sequence conservation across species. The use of conservation as an indication of function is based on the hypothesis that functional sequences in the genome will be less tolerant of variation than regions that are non-functional. Accordingly, if one examines a speciation event at a hypothetical single locus it is observed that at first (time=0) sequence identity is 100 percent, but over extensive evolutionary time random mutations will arise across the genome during replication. Those that decrease the fitness of the organism are not tolerated and will be lost while those that have no effect on fitness will be allowed to accumulate. This results in varying amounts of sequence conservation across the genome from which one can predict functional elements. Support for this hypothesis is found in the fact that exons of genes show very high conservation while introns tend to show lower conservation (Miller et al., 2004). Therefore, noncoding regions with high interspecies conservation are predicted to be functional and can be selectively cloned from the genome for further functional analysis.

Our lab has used screening regions of interest for conservation for more than 10 years and my discussion on the identification of enhancers at *LMX1A* and *LMX1B* applied this strategy directly (Grice et al., 2005; Fisher et al., 2006; Antonellis et al., 2008; Burzysnski and Reed et al., 2013). However, there are a few drawbacks, the most significant being that it is a locus-based analysis and extrapolation to the whole genome would be a monumental and likely prohibitive task. We have also previously shown that analysis of conserved regions alone misses a significant portion of functional noncoding sequences (McGaughey et al., 2008; McGaughey et al., 2009). Additionally, the presence of conservation gives no information regarding the function of a region or the gene upon which it may act. For these

reasons, we are constantly investigating new technologies and molecular methods that will allow for genome-wide identification of enhancers within a specific cell type.

1.4 Molecular methods to identify enhancers

Chromatin immunoprecipitation (ChIP) for transcription factors, histone modifications and/or transcriptional co-activators has become a commonly used strategy to show that a particular DNA sequence interacts with known proteins (O'Neill and Turner, 1995). ChIP is based on the physical interaction of proteins with specific DNA sequences. Cells are generally cross-linked to maintain the *in vivo* interactions and then cells are lysed, DNA fragmented, and antibodies are used to enrich for the protein of interest, and thus the DNA that it has bound (O'Neill and Turner, 1996). Initially these methods were limited to predefined loci, however as technology has improved, first with chips to employ microarray hybridization of immunoprecipitated DNA (ChIP-chip; Horak et al., 2002) then with high throughput next-generation sequencing (ChIP-seq; Robertson et al., 2007) they have become scalable to analysis of the whole-genome.

It was first shown in 2007 by ChIP-chip that a specific histone modification, H3K4me1, is associated with enhancers that are characterized by binding of the transcriptional co-activator EP300 (Heintzman et al., 2007). ChIP-seq for EP300 as a marker of enhancers was first used in 2009 to identify enhancers genome-wide in the forebrain, midbrain and limb of embryonic mice (Visel et al., 2009). Further ChIP-seq experiments for histone modifications found enrichment for both H3K4me1 and H3K27ac at putative enhancers defined by the presence of EP300 (Rada-Iglesias et al., 2011). As the cost of high throughput sequencing has decreased in recent years its frequency of use has eclipsed that of sequence conservation alone.

DNase-seq is another genome-scalable method that allows for the broad capture of DNA in euchromatin, a prediction of functionality, and has also been used to create catalogs of putative enhancers (Song and Crawford, 2010; Mercer et al., 2013). Molecular methods for the discovery of enhancers, like ChIP-seq, are desirable because they are unbiased by the conservation status or location of a putative enhancer and therefore provide a broad look at the noncoding genomic regions predicted to be functional within a particular cell type. This makes them an ideal starting point for discovering catalogs of enhancers and characterizing the vocabularies associated with them in specific neuronal populations

1.5 Dopaminergic neurons and human disease

Dopaminergic (DA) neurons are characterized by the use of dopamine as a signaling molecule. The most dense population of DA neurons is found in the *substantia nigra* (SN), in the ventral midbrain, where they play a key role in voluntary movement. Smaller, more sparse, populations of DA neurons exist throughout the CNS in the hypothalamus, olfactory bulb, retinal amacrine cells, striatum and the ventral tegmental area (VTA) where they are hypothesized to influence diverse processes such as satiety, scent, reward pathways, mood and other cognitive behaviors (Prakash and Wurst, 2006). The variety of functions makes DA neurons a very important contributor to many aspects of human health and disease.

Degeneration of DA neurons of the SN leads to the akinesia, hypokinesia and/or bradykinesia that characterize Parkinson's Disease (PD), while disruptions of the pathways encompassing DA neurons in the striatum and VTA have been implicated in addiction, depression, bipolar disorder (BP) and schizophrenia (SZ; Robinson et al., 1993; Tzschentke et al., 2000; Meyer-Lindberg et al., 2002). PD and SZ are among the world's most common neurological disorders, each having a population incidence of ≥1% (Savitt et al., 2006).

Mutations in at least 8 genes have been shown to cause Mendelian forms of Parkinson's and Genome Wide Association Studies (GWAS) have identified more than 19 other loci, some affecting the same genes that are mutated in Mendelian forms (Table 1-1; Singleton et al., 2013). Our incomplete understanding of the molecular genetics controlling DA neuronal differentiation, survival and competence is a major obstacle to novel prophylactic and preventative strategies. The studies that follow aim to begin the process of identifying and characterizing the transcriptional networks underlying the development and function of this clinically relevant neuronal subtype.

1.6 Functional assays in zebrafish

Upon identification of a putative neuronal enhancer we must investigate if, when and where the sequence drives expression. Zebrafish are a good model system for this analysis for a number of reasons. First, embryos are transparent, undergo external development and develop a fully functional nervous system after 4-5 days of development (Kuwada, 1995; Budick and O'Malley, 2000). Second, both the genome (Gates et al., 1999; Barbazuk et al., 2000) and brain development and structure are largely conserved between zebrafish and mammals (Figure 1-2; Mueller and Wullimann, 2003; Guo, 2004; Mueller et al., 2006; Aizawa et al., 2011). Zebrafish contain all of the major cell types found in the CNS (Westerfield et al., 1986; Koulen et al., 2000; Kawai et al., 2001; Yoshida et al., 2005; Avila et al., 2007; Peri and Nusslein-Volhard, 2008), and share many of the same neuroanatomical (Rink and Wullimann 2002; Mueller and Wullimann 2009) and neurochemical pathways with humans (Holzschuh et al., 2003; Mueller et al., 2006; Filippi et al., 2010; Filippi et al., 2014). Additionally, zebrafish have a relatively short time to sexual maturity (~ 3 months) and each female can produce hundreds of embryos each week, making it easier to obtain

large numbers for analysis than with transgenic mice. For these reasons, zebrafish reporter assays allow for a quick and cost effective analysis of putative neuronal enhancers.

Reporter assays in zebrafish are very straightforward and are done by first designing PCR primers to the genomic region of interest. The sequence is then PCR amplified and cloned into a vector upstream of a minimal promoter linked to a fluorescent reporter. The minimal reporter alone does not drive reporter expression but in the presence of a sequence that acts as an enhancer it will drive expression in the embryo wherever that sequence is active. Our reporter construct is also flanked by tol2 sequences that allow it to randomly integrate into the zebrafish genome when co-injected with tol2 transposase into fertilized zebrafish embryos (Fisher et al., 2006a and b). Fluorescent readout can be visualized in the developing embryos to determine when and where the sequence drives expression. One limitation of these assays is the possibility of interaction between the minimal promoter and the genomic DNA flanking the site of integration, dubbed position effects (Fisher et al., 2006b). These effects can be overcome by the analysis of many transgenic embryos to identify overlapping regions of expression as the true enhancer readout. *In vivo* reporter assays are especially useful when examining specific neuron populations that may not have an appropriate cell line available for luciferase activity analysis, such as dopaminergic neurons.

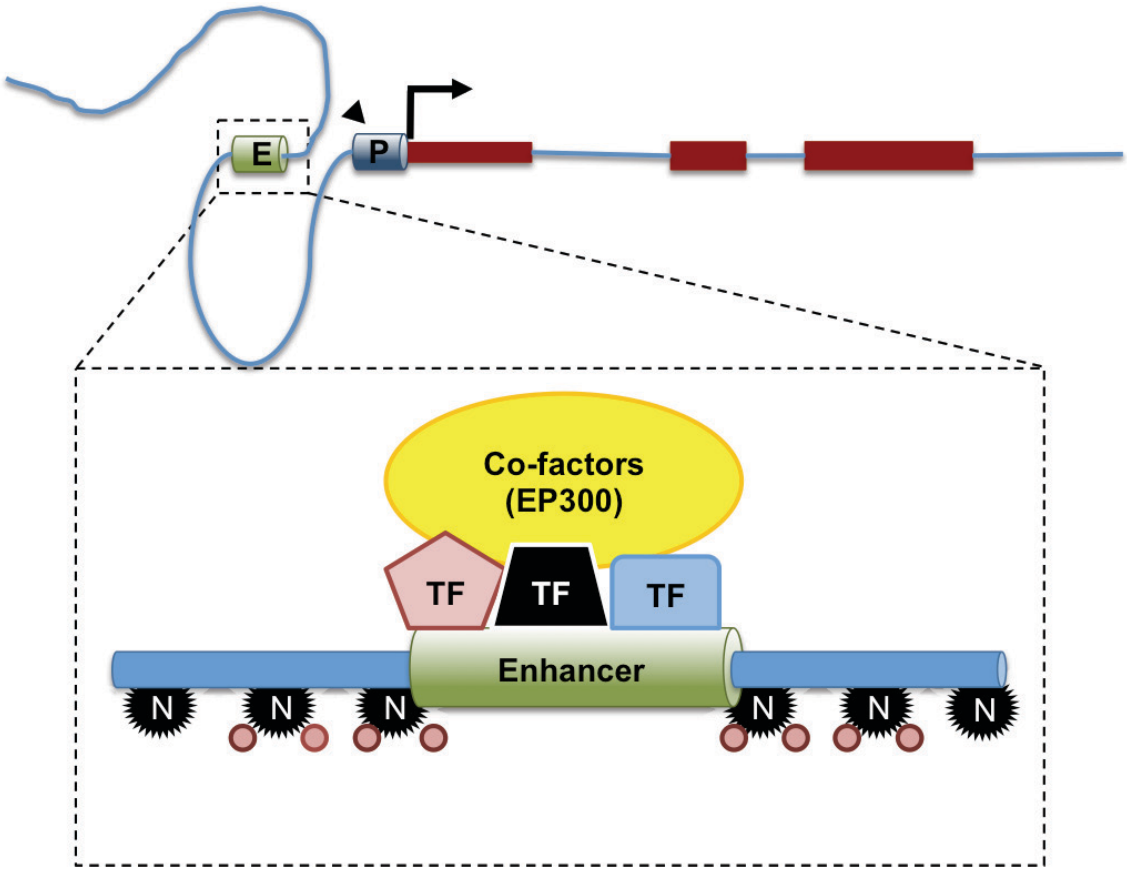
1.7 Tables: Chapter 1

Table 1-1. Genes with coding mutaions linked to familial parkinsonian phenotypes.

Gene	Location	Heritability	OMIM #
SNCA	4q22.1	AD	168601, 605543
LRRK2	12q12	AD	607060
VPS35	16q11.2	AD	614203
EIF4G1	3q27.1	AD	614251
DNAJC13	1p31.3	AD	614334
PARK2	6q26	AR	600116
PINK1	1p36.12	AR	605090
DJ-1	1p36.23	AR	606324
DNAJC6	1p31.3	AR	615528
ATP13A2	1p36.13	AR	606693
FBXO7	22q12.3	AR	260300
PLA2G6	22q13.1	AR	612953
ATP6AP2	Xp11.4	X-linked recessive	300911
SYNJ1	21q22.11	AR	615530

1.8 Figures: Chapter 1

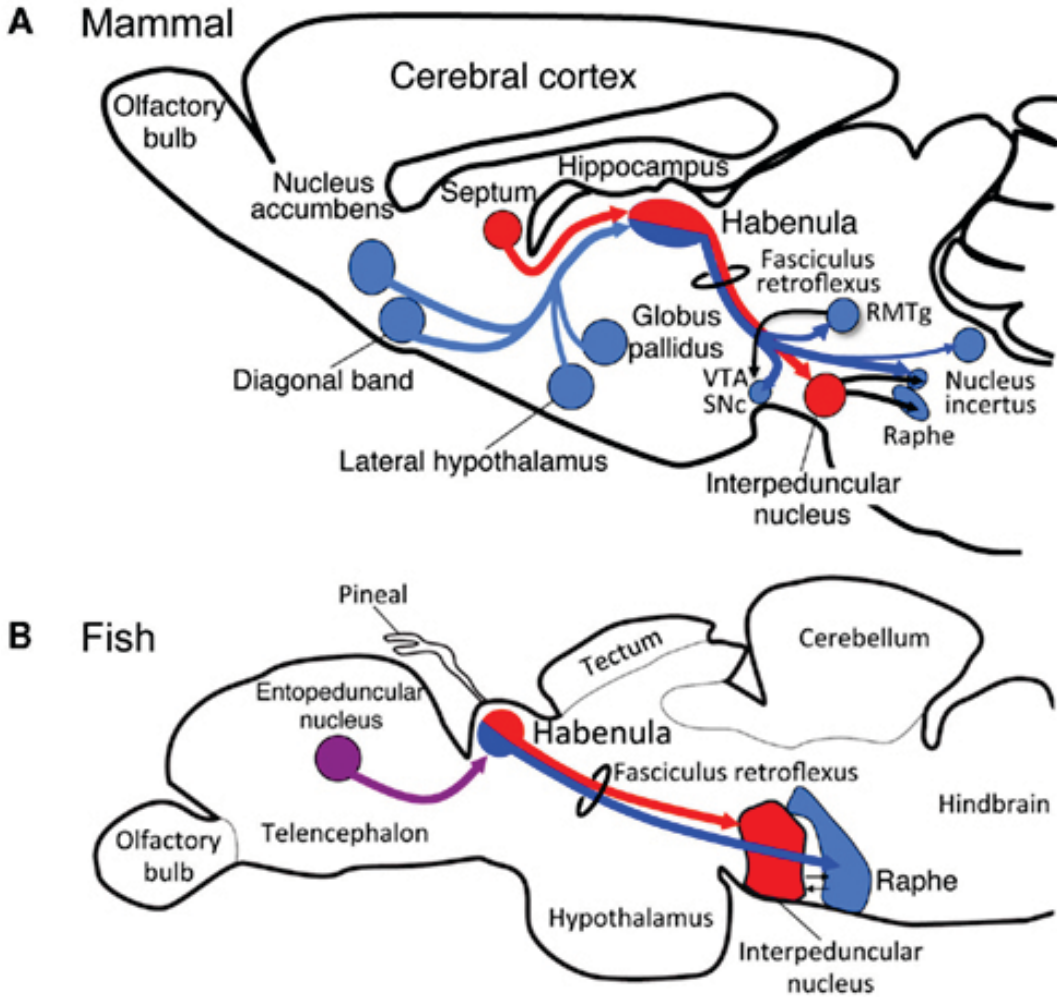
Figure 1-1. Model of enhancer function.



Nucleosomes flanking enhancers are enriched for H3K4me1 (red circles). Transcription factors bind the enhancer and recruit co-factors such as EP300 and the mediator complex to interact with the promoter and in turn recruit RNA polymerase. Adapted from Noonan and McCallion, 2010 and Gorkin, 2013.

Abbreviations: E – Enhancer; N – Nucleosome; P – Promoter; TF – Transcription factor.

Figure 1-2. Zebrafish brain organization is largely conserved with mammals.



Representation of saggital sections from A) rat brain and B) zebrafish brain. Although the structure and organization are slightly different, functional regions and cicuitry are retained in zebrafish. Red and blue lines show examples of conserved circuitry in the habenular pathways. Adapted from Aizawa et al. 2011.

Abbreviations: SNc – Substantia nigra pars compacta, RMTg – rostromedial tegmental nucleus, VTA – Ventral tegmental area.

CHAPTER 2

USE OF ESTABLISHED METHODS OF SEQUENCE CONSERVATION AND REPORTER ASSAYS FOR ENHANCER DISCOVERY

2.1 Introduction

In this chapter we examine sequences predicted to regulate the expression of the lim homeodomain containing transcription factors, *LMX1A* and *LMX1B*. These genes play critical roles in the development of the central nervous system, specifically in neural tube regionalization, hindbrain development, the development of axonal projections and differentiation into dopaminergic and serotonergic neuronal subtypes (Hobert and Westphal, 2000; Shirasaki and Pfaff, 2002; Dai et al., 2009; Andersson et al., 2006). *Lmx1b* has been shown in mice to be required for normal mesencephalic dopaminergic neuron development, serotonergic neuron development, podocyte differentiation and numerous cases of inactivating *LMX1B* mutations have been found to be responsible for human nail patella syndrome (Smidt et al., 2000, Zhao et al., 2006, Miner et al., 2002, Vollrath et al., 1998). Similarly, *Lmx1a* mutations were initially described in mouse neurological mutant *dreher*, which displays defects in cerebellar, hippocampal, and cortical development, as well as hindbrain roof plate malformations, short tail and deafness consistent with the patterns of its embryonic expression (Millonig et al., 2000; Failli et al., 2002).

Despite the developmental importance of these genes their regulation is not well understood. When examining specific loci evolutionary sequence conservation provides a powerful tool for the identification of functional sequences, and although conservation alone is unable to discern the biological roles of sequences, through functional analyses, one can discover regulatory elements. In order to begin to investigate the transcriptional regulation at

LMX1A and *LMX1B* we have identified 71 conserved noncoding regions flanking these loci. We first characterize the expression patterns of the homologous zebrafish genes (*lmx1a, lmx1a-like, lmx1ba,* and *lmx1bb*) in order to describe their overlap with the patterns driven by putative enhancers flanking their human counterparts. Next we test the ability of these noncoding sequences to drive expression in zebrafish reporter assays. This reporter analysis identifies 22 previously unknown enhancers, 17 flanking *LMX1A* and 5 flanking *LMX1B*, which drive expression in the CNS and may be important in regulating gene transcription during the development and maintenance of dopaminergic and serotonergic neuron populations.

2.2 Results

2.2.1 Evolutionary conservation facilitates identification of zebrafish lmx1a and lmx1b genes

In order to effectively evaluate the expression patterns of putative enhancers flanking *LMX1A* and *LMX1B* in zebrafish one must first characterize the expression of the homologous zebrafish genes. Therefore, we first set out to identify homologs of *LMX1A* and *LMX1B* in zebrafish. Approximately 30% of the gene content of *Danio rerio* is duplicated as a result of an ancient genomic duplication event in the teleost fish lineage (Amores et al., 1998). Due to this duplication, the zebrafish genome contains two identified *LMX1B* paralogs (*lmx1ba* and *lmx1bb*). However, only one *LMX1A* paralog (*lmx1a*) had been identified in the zebrafish genome at the time of these experiments. We performed a BLASTP query of the zebrafish peptide database in GenBank using the human *LMX1A* RNA sequence (NM_001174069.1) and identified another potential paralog previously annotated with 'predicted' status (LIM homeobox transcription factor 1-alpha-like, XP_001922131.3).

LMX1A is 66% identical to LMX1B at the amino acid level, and is 58% and 59% identical to zebrafish Lmx1a and Lmx1a-like, respectively. The LMX1B homologs are more similar with Lmx1ba having 72% protein sequence identity and Lmx1bb being 82% identical to the human LMX1B protein sequence (NP_001167617.1). Figure 2-1 provides a phylogram illustrating the similarity among the amino acid sequences that encode LMX1A, LMX1B and their zebrafish paralogs. The paralogs of *LMX1A* cluster together, but in a distinct node from their human counterpart. By contrast, *LMX1B* shares a common node with its zebrafish paralogs.

2.2.2 Zebrafish lmx1 genes are expressed throughout the central nervous system

After identification of zebrafish *lmx1* homologs we performed whole mount *in situ* hybridizations (ISH) to examine expression of the endogenous zebrafish *lmx1* genes. ISH allows us to document the spatial and temporal expression patterns of the endogenous *lmx1a*, *lmx1a-like*, *lmx1ba* and *lmx1bb* mRNAs, and to determine the similarity to the published expression of their mammalian orthologs in mice. Aspects of the early developmental expression of *lmx1ba* and *lmx1bb* have been previously described, however we repeated their ISH analysis to ease comparisons with expression patterns of *lmx1a* and *lmx1a-like* (O'Hara et al., 2005; Cheng et al., 2007; Elsen et al., 2008; McMahon et al. 2009; Filippi et al., 2010)..

Transcript from *lmx1a* was seen diffusely throughout the brain, including the diencephalon and telencephalon, at both 48 hours post fertilization (hpf) and 72 hpf (Figure 2-2 A-D). We detected a more distinct and localized signal in the ventral diencephalon, raphe nuclei and otic vesicles at both time points, and at 72 hpf saw specific labeling of the cranial ganglia (Figure 2-2 A-D). In contrast, *lmx1a-like* expression was regionally restricted, with

distinct labeling of the epiphysis, ventral diencephalon, rhombic lip, and raphe nuclei, closely resembling expression of *LMX1B* paralogs (Figure 2-2 E-L). We also detected expression of *lmx1a-like* in the anterio-dorso-lateral hindbrain and in the ventro-midline, corresponding with the cerebellar rhombic lip and serotonergic raphe nuclei, respectively (Figure 2-2 E-H). Both *lmx1a* and *lmx1a-like* appeared to be more highly expressed in the anterior raphe nuclei at 48 hpf (Figure 2-2 A, B, E, and F), while the *lmx1b* paralogs were seen approximately equally in both raphe nuclei populations (Figure 2-2 I, J, M, and L). These data are consistent with both *Lmx1a* and *Lmx1b* mammalian counterparts, which are expressed in the developing cerebellum and serotonergic neurons. *lmx1a-like* has very little dorsal hindbrain expression at 48 hpf but by 72 hpf transcripts are detected strongly in the posterior dorsal hindbrain (Figure 2-2 G and H). This pattern is unique to *lmx1a-like* while some expression domains overlap expression of *lmx1ba* and *lmx1bb* in the ventral diencephalon, rhombic lip, serotonergic raphe nuclei and faintly in the otic vesicles (Figure 2-2 I, J, M and N; Cheng et al., 2007; Filippi et al., 2010).

The patterns of expression are very similar between *lmx1ba* and *lmx1bb* with shared domains in the ventral diencephalon, raphe nuclei, rhombic lip, and dorsal hindbrain, as well as the amacrine neurons of the retina at 72 hpf (Figure 2-2 I-P). *lmx1bb* and, to a lesser extent, *lmx1ba* show additional expression in the dorsal diencephalon that is not seen for the *lmx1a* transcripts (Figure 2-2 I-P). Overall *lmx1bb* shows broader domains of expression than *lmx1ba* throughout the CNS, however *lmx1ba* transcript is also unexpectedly detected in the heart (Figure 2-2 I-L). Notably, strong expression is seen for all transcripts in the ventral diencephalon, the area where main clusters of dopaminergic (DA) neurons are formed, through 72 hpf consistent with their roles in the induction and specification of midbrain DA neurons (Andersson et al., 2006; Mishima et al., 2009; Yan et al., 2011).

2.2.3 Selection of noncoding sequences at human LMX1A and LMX1B loci

The human *LMX1A* gene is made up of seven exons, encompassing 154 kb on chromosome 1q24 and is flanked by *PBX1* and *RXRG* (Figure 2-3 A). *LMX1B* includes 8 exons that cover 112 kb of chromosome 9q33.3 and is flanked by *FAM125B* and *ZBTB43* (Figure 2-3 B). Candidate sequences for functional analysis were selected from the intervals between their respective flanking genes, therefore providing *LMX1A* and *LMX1B* regions of 554 kb and 298 kb, respectively. Although this search space was not exhaustively explored, we prioritized conserved noncoding sequences for assay using Genomic Evolutionary Rate Profiling (GERP; Cooper et al., 2005). We successfully PCR amplified 71 noncoding DNA sequence intervals (see methods; 43 sequences at *LMX1A* and 28 at *LMX1B*). These putative enhancers were cloned into pGWcfos: EGFP and injected into fertilized zebrafish embryos (Fisher et al., 2006 a and b). All embryos showing mosaic EGFP expression in the CNS (33 *LMX1A* (77%) and 14 *LMX1B* (50%)) were separated and raised for germline transmission analysis.

2.2.4 Tested sequences display LMX1A and LMX1B-appropriate EGFP activities

Of the assayed sequences, 37 displayed reporter expression in the CNS upon passage through the germline. We identified two or more founders with concordant expression from 22/37 (59%) conserved noncoding regions (*LMX1A*, n=17 and *LMX1B*, n=5). In general, the lines for which we were unable to identify multiple founders suffered from poor survival and fecundity. Those elements were therefore most often excluded, not because of divergent expression patterns but due to the inability to obtain a sufficient number of fertilized embryos for screening.

All 22 sequences in which multiple founders were identified displayed spatial control in the CNS that overlapped, at least in part, with the endogenous expression patterns described above (Table 2-1; Failli et al. 2002). This includes directing reporter expression within discrete regions of the diencephalon, telencephalon and hindbrain. We also identified enhancers at *LMX1A* that display regulatory control of reporter expression resembling the more diffuse expression of *lmx1a* in the CNS (LMX1A_-1.59; Table 2-1). Consistent with their neuronal activity in our synthetic assay, the majority of endogenous sequence intervals corresponding to our assayed sequences also displayed enrichment for the enhancer associated H3K4me1 mark in two lines of cultured neurospheres derived from human neuronal cells Neurosphere Cultured Cells, Ganglionic Eminence Derived (NGED) and Neurosphere Cultured Cells, Cortex Derived (NCD) (Figure 2-3 A-C; Heintzman et al. 2007; Bernstein et al. 2010). Indeed, despite lacking a positive GERP alignment score, sequence LMX1A_36.74 displayed strong H3K4me1 binding in both NGED and NCD cells and was validated in our zebrafish assay (Figure 2-4 E).

2.2.5 *Identification of LMX1A enhancers with telencephalic and diencephalic regulatory control*

Consistent with the endogenous expression of the mouse *Lmx1a* mammalian ortholog (Faili et al., 2002), we identified many *LMX1A* enhancers displaying overlapping control in the telencephalon (Figure 2-4 A-C, D, E; and Table 2-1). Telencephalic expression displayed by these sequences is consistent with the function of LMX1 genes in cortical hem development (Figure 2-4 A-E; Chizhikov et al., 2010; Adams et al., 2000; Guo et al., 2007). Telencephalic expression is also evident for the zebrafish *lmx1a*, however it is diffuse and and not at significant levels (Figure 2-2 A-D). This observation may thus reflect mammalian

(*LMX1A*/*Lmx1a*) control alone in this structure. We identified multiple additional sequences at *LMX1A* that direct expression in the diencephalon (Figure 2-4 A, D-F; and Table 2-1 e.g. LMX1A_238.85, LMX1A_-36.74 and LMX1A_9.65). These populations may include portions of the catecholaminergic diencephalic cluster, consistent with the established role of *Lmx1a* in mouse catecholaminergic neurogenesis (Yan et al., 2011).

2.2.6 Many identified LMX1A and LMX1B enhancers display regulatory control at the midbrain/hindbrain boundary and in the hindbrain

Multiple REs from both loci are able to drive expression in the midbrain-hindbrain boundary region that includes the IsO and anterior cerebellum (Figure 2-5 and Table 2-1; e.g. LMX1A_-36.74, LMX1A_41.10, LMX1A_479.86, LMX1A_135.15 and LMX1B_-79.84). These data corroborate with the role of mammalian LMX1 genes in IsO and cerebellum development and function (Adams et al., 2000; Guo et al., 2007). Additionally, many assayed sequences directed expression in the hindbrain (Figure 2-5 and Table 2-1; e.g. LMX1A_475.57, LMX1A_238.85, LMX1A_5.34, LMX1A_-1.59 and LMX1B_28.46), including the roof plate (LMX1A_102.3) and the spinal cord (Table 2-1; e.g. LMX1A_-1.59, LMX1A_238.85 and LMX1B_28.46) consistent with the endogenous expression of their corresponding zebrafish paralogous transcripts.

2.2.7 LMX1 enhancers display regulatory control in peripheral neuronal as well as non-neuronal cell populations

We identified several *LMX1A* and *LMX1B* sequences that direct expression in the otic vesicle (Figure 2-4 B, and Table 2-1, e.g. LMX1A_-36.74, LMX1A_92.33 and LMX1A_117.21, LMX1B_-93.21), consistent with the expression of Lmx1a (Failli et al.,

2002). Furthermore, mice deficient in Lmx1a display abnormal ear development and deafness (Millonig et al., 2000; Huang et al., 2008). Multiple lines display reporter expression in PNS structures, like motor neurons (Figure S2 and Table 2-1; LMX1A_238.85) or sympathetic chain (Figure S2 and Table 2-1; e.g. LMX1A_475.57). In contrast to the well-characterized neuronal roles of *LMX1A* and *LMX1B*, many identified enhancers also drive reporter expression in non-neuronal tissues such as the branchial arches, (Table 2-1, Figure 2-4 A-C, Figure S2) which are also documented sites of expression in the mouse (Chen et al., 2003; Failli et al., 2002). We also identified a single *LMX1B* enhancer that displays expression in the heart (LMX1B_-93.21). Although unexpected, this is consistent with our ISH for endogenous expression of *lmx1ba*. The biological significance of *lmx1ba* has not yet been determined but may, in part, correspond to the aortic arch neurons where expression of the catecholaminergic marker tyrosine hydroxylase has been previously reported (Wen et al., 2008).

2.3 Conclusions

In order to better understand the regulatory landscape of *LMX1A* and *LMX1B* we undertook a functional study of conserved, noncoding sequences (putative REs) at these loci, using zebrafish transgenesis. We first established the identity of two zebrafish paralogs for each human LMX1 gene. We then demonstrated that their endogenous expression closely resembles the previously characterized expression of their mammalian counterparts, including expression in the areas of presumptive catecholaminergic neurons, cerebellum, raphe nuclei and otic vesicles (Figure 2-2). Next, we used comparative sequence analyses to identify conserved, noncoding sequences at the human LMX1A and LMX1B loci, selecting 71 putative REs for functional evaluation. Of these putative enhancers, 45 directed CNS

reporter expression in the CNS of mosaic zebrafish embryos. We further described the reporter expression of 22 sequences in stable transgenic lines (LMX1A, n=17; LMX1B, n=5). All 22 display consistent CNS enhancer function (n≥2 independent founders) that overlaps, at least in part, with the endogenous transcripts. The majority of these sequences display enrichment for H3K4me1, a modification known to be enriched at enhancers, in cultured neurospheres (Figure 2-3; Heintzman et al., 2007; Bernstein et al. 2010), consistent with their neuronal activity in our synthetic assay, and providing evidence supporting their likely cis-regulatory role in their endogenous context. Instances where we do not detect reporter expression most often reflect failure to identify more than one founder transgenic line and not inconsistencies among multiple lines for a single construct.

The diencephalon, telencephalon and midbrain-hindbrain boundary were among the most common structures marked by reporter expression for REs identified at both loci (Figure 2-4; Table 2-1; Appendix 1). Many enhancers similarly directed broad expression in the midbrain (Figure 2-4; Table 2-1; Appendix 1) and more discrete expression in the hindbrain e.g. in single rhombomeres, area postrema (Figure 2-5; Table 2-1; Appendix 1, LMX1B_-79.84) or hindbrain roof plate (Figure 2-4; Table 2-1; Appendix 1, LMX1A_102.3). These sites of expression overlap known domains of *Lmx1a* and *Lmx1b* expression in mammals and teleosts. The expression directed in the midbrain-hindbrain boundary, cerebellum and posterior rhombomeres is consistent with the important function of both *LMX1A* and *LMX1B* in the development of the cerebellum rhombic lip and hindbrain roof plate (Mishima et al., 2009; Chizhikov et al., 2010). Furthermore, *LMX1B* is known to be instrumental for proper functioning of the IsO [(Adams et al., 2000; Guo et al., 2007). We hypothesize that the forebrain expression patterns of some *LMX1A* enhancers reflect the gene's role in early development of cortical hem in mammals.

Many sequences also direct expression in the PNS and some non-neuronal tissues, consistent with the endogenous expression of these TFs and their roles in the differentiation and maintenance of a range of populations. The most common non-neuronal site of reporter expression was found in the otic vesicles, which is consistent with previously published *LMX1A/B* biology (Millonig et al., 2000; Failli et al., 2002; Huang et al., 2008). Interestingly we also identified some enhancers that drive expression in the heart, which may correspond to peripheral neuronal populations (Wen et al., 2008).

Transgenic assays provide an approximation of how a regulatory sequence can behave in a model system and may not capture every nuance that the corresponding sequence may display in context. Therefore, their correspondence to the spatial expression of lmx1 genes does not definitively demonstrate their control (exclusive or shared with neighbors) of these genes. In particular, we recognize that aspects of the CNS regulatory control displayed by enhancers isolated near *LMX1A* and *LMX1B* may also be considered consistent with expression of the flanking genes. In particular, transcript from the zebrafish *pbx1a* paralog is seen in many domains that also show expression of *lmx1a/lmx1a-like* (Thisse et al., 2004), including discrete expression in the telencephalon. Thus firm conclusions regarding enhancer driven reporter expression and their direct relation to LMX1-expressing neuronal populations or those of their flanking genes will require additional experimental determination of possible physical interaction between enhancer and one or more cognate promoter. Despite these caveats, these assays can and do provide significant insight into the biological relevance of assayed sequences.

We demonstrate how a range of available data types may be integrated in the exploration of the genomic information content of sequence encompassing two critical human genes. This work is the first to describe the endogenous expression patterns of all

zebrafish *LMX1A* paralogs; it identifies 22 previously unknown enhancers and sheds light on previously unknown transcriptional regulatory landscape at the *LMX1A* and *LMX1B* loci. If one accounts for the presence of additional conserved and/or histone-marked sequences in the genomic intervals under consideration, these enhancers may represent only a fraction of the conserved noncoding elements at these loci. We hypothesize that many enhancers may be required in combination to orchestrate regulatory control of these genes. The multitude of neuronal and non-neuronal populations marked by these conserved noncoding sequences may highlight additional complexity in this regulatory control or reflect position effects. These data reinforce the value of targeted screens in the analysis of human disease loci and integrating comparative sequence analysis, chromatin modifications and functional validation using zebrafish transgenesis in the identification of transcriptional regulatory sequences.

2.4 Methods

2.4.1 Fish maintenance

Zebrafish were kept and bred under standard conditions at 28.5ºC (Westerfield, 2000). Embryos were staged and fixed at 48 and 72 hpf using 4% paraformaldehyde (PFA) in phosphate-buffered saline (PBS; pH 7.2) as described elsewhere (Kimmel et al., 1995). To better visualize in situ hybridization and EGFP reporter results, embryos were grown in 0.2 mM 1-phenyl-2-thiourea (PTU; Sigma) to inhibit pigment formation (Westerfield, 2000).

2.4.2 Whole mount in situ hybridization

Digoxigenin labeled riboprobes complementary to *lmx1a*, *lmx1a-like*, *lmx1ba* or *lmx1bb* mRNAs were generated by linearization of pCR II TOPO TA vectors containing partial ORFs of the genes (for probe sequences see Table 2-1). Plasmids were linearized with

EcoRV (New England Biolabs) and subsequently labeled riboprobes were transcribed using SP6 polymerase and the DIG RNA Labeling Kit (T7/SP6) (Roche). Probes were synthesized for 2 hours at 37ºC, followed by the addition of 1 μl of RNAse free DNAse I for DNA template digestion. Subsequently, probes were purified using SigmaSpin columns (Sigma-Aldrich). Whole mount in situ hybridization reactions were performed using 1:4000 dilutions of rioboprobes at 70˚C as previously described (Thisse et al., 1993 and 2003). Probe sequences were selected to avoid cross-hybridization with lmx1 family members and unrelated transcripts by using pairwise alignment of lmx1 transcripts to find unique stretches of mRNA. Sequences were aligned using Clustal Omega (http://www.ebi.ac.uk/Tools/msa/clustalo/).

2.4.3 Selection and amplification of human noncoding sequences

To select regions to test for potential enhancer activity, genomic intervals encompassing *LMX1A* and *LMX1B* loci were considered up to the neighboring genes, set as boundaries, (*LMX1A*, chr1: 163,082,934 - 163,636,974 bp; *LMX1B*, chr9: 128,309,140 – 128,607,182 bp). This study is not intended to be exhaustive and the genomic intervals encompassing these genes are very large. Therefore, sequences were prioritized for selection based upon proximity to the LMX1 loci and conservation.

Using the Galaxy computational interface (Goecks et al., 2010) and UCSC genome browser, we chose conserved noncoding vertebrate elements with positive GERP alignment scores (Cooper et al., 2005). The GERP algorithm identifies constrained sequences in genomic alignments by determining whether a scarcity of substitutions exists at each point in an alignment relative to what is expected of the neutral rate of evolution. Selected intervals positioned less than 500 bp apart were merged into single amplicons. DNA region

coordinates and primer sequences used for amplification are listed in Appendix 2. Sequences were also selected to avoid clustering and are distributed across these loci (Figure 2-2 A). Amplicons were PCR amplified from human genomic DNA, TA cloned into pCR8 (Invitrogen) and then cloned using the Gateway system (Invitrogen) into pGW_cfosEGFP as previously described (Fisher et al., 2006a and b).

2.4.4 Embryo injection and analysis

EGFP reporter constructs were injected into AB background G0 embryos (n≥200) at the one to two cell stage with tol2 transposase as previously described (Fisher et al., 2006a and b). Injected embryos were evaluated for EGFP expression between 24 and 72 hpf. As negative controls EGFP reporter constructs containing only the cFos promoter were injected. Nonspecific expression from the cfos minimal promoter is occasionally observed in the myotome and no other nonspecific expression was detected. Embryos showing consistent EGFP expression were selected and raised for further analysis when signal was observed in ≥10% of injected embryos. Mosaic fish were subsequently crossed to identify those constructs that passed through the germline transmission, better facilitating spatial evaluation of corresponding EGFP expression. Embryos were imaged using a Carl Zeiss Lumar V12 Stereo microscope with AxioVision software (version 4.5).

2.4.5 Immunocytochemistry

Embryos were anesthetized with tricaine (10 μg/ml) in embryo medium (Westerfield, 2000) and fixed in 4% PFA in phosphate-buffered saline (PBS; pH 7.2) for 2 hours. They were then rinsed four times in PBST (PBS/0.1% Triton X-100), incubated in Proteinase K (Roche) for 1h at room temperature, washed 5 × 5 minutes in PBST, and incubated for 2

hours in blocking solution (10% goat serum, 1% bovine serum albumin (BSA), in PBST). Embryos were then incubated overnight at room temperature in primary antibody (anti-GFP, Invitrogen 1:2000), rinsed 6 × 45 minutes in PBST with 1% goat serum, and incubated overnight at room temperature in secondary antibody (Alexa-Fluor, 488, Invitrogen 1:1000). They were then rinsed 5 × 10 minutes in PBST and transferred to 50% glycerol in PBS for imaging.

2.5 Tables: Chapter 2

Table 2-1. Primer sequences used for generation of RNA in situ probe fragments.

Transcript name	Primer sequence	RefSeq alignment	Length (bp)
lmx1a-F	ACTCTCTGGATAATGATGTGCCA	NM_001025498	201
lmx1a-R	ATTGCGGAGAAAGCAGGTGT	4 - 204	
lmx1a-like-F	AACACTGAAGTTTGGCTTTTGGA	XM_001922096	260
lmx1a-like-R	GATGGGGGACTCGCAGC	51 - 310	
lmx1ba-F	CTCTCGACAGATCCGGCTG	NM_001025168	231
lmx1ba-R	TCAATTTTGATTCCGTCCAGCA	104 - 334	
lmx1bb-F	GCCGAATGGCGCACTAATTT	NM_001025167	296
lmx1bb-R	GCAGCGGATGATCTTCGATTTT	23 - 318	

Primers designed against zebrafish LMX1 cDNA paralogs and the RefSeq database alignments used for each.

Table 2-2. Systematic annotation of LMX1A and LMX1B enhancer activity in zebrafish body structures.

Enhancer	Tel	Dien	Mesen	Rhom	SC	Mb/Hb	PNS	Other
LMX1A -36.74	+	+	+	+		++	+	
LMX1A -1.59	++	++	++	++	++	+	+	
LMX1A 5.34				+	++	+		NC
LMX1A 9.65	+	+		+	+			OV
LMX1A 11.50	+	+	+	+		+	+	
LMX1A 32.09		+	+	+	+			Pn, L
LMX1A 67.70	++	+			+		+	NC
LMX1A 71.89	+	+	+	+			+	C, F, H, OV, R
LMX1A 92.33	+				+			C, F, OV
LMX1A 102.30	++	+	+	++	+	+		
LMX1A 117.21	+	+	+	+	+	+	+	C, H
LMX1A 135.15	+	+	+	+	+	++	+	
LMX1A 187.31	+	+	+	+	+		++	
LMX1A 238.85		+	+	++	++		++	
LMX1A 296.85	++	+		++	+			
LMX1A 475.56	+	+	+	+	+		+	OV
LMX1A 479.56	+		+	+		++		C
LMX1B -93.21				+			+	C, H
LMX1B -79.84			++	+	+	++	+	
LMX1B 12.07	+	+	+	+	+	+	+	U
LMX1B 28.46	+	+	+	+	+	+		L
LMX1B 59.40	+	+	+	+	+		+	

Abbreviations: C – Cartilage, Dien – Diencephalon, F – Fins, H – Heart, L – Lens, Mesen – Mesencephalon, Mb/Hb – Midbrain/Hindbrain, NC – Notochord, OV – Otic Vesicle, PNS – Peripheral Nervous System, Pn – Pronephros, R – Retina, Rhom – Rhombencephalon, SC – Spinal Cord, Tel – Telecephalon, U – Ubiquitous.

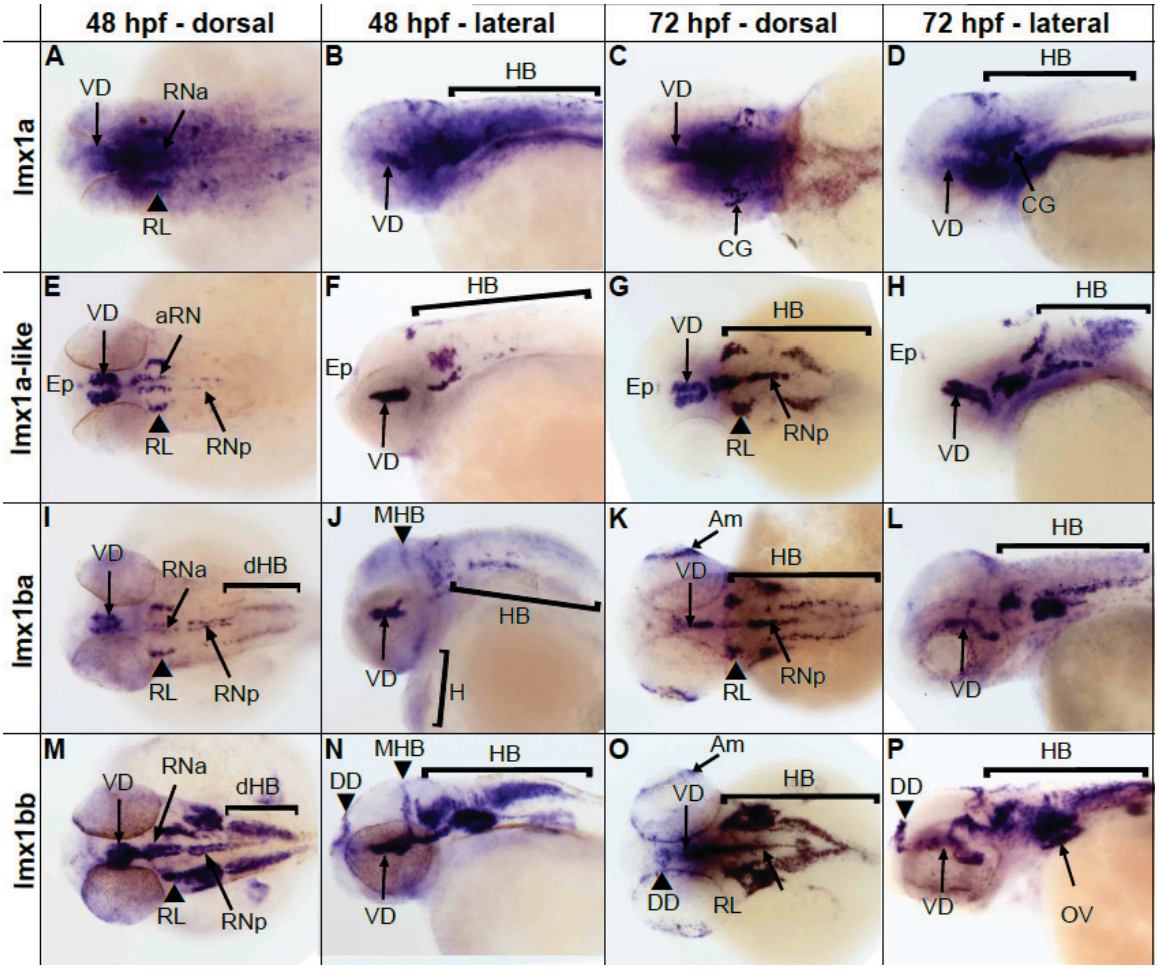
+, Weak; ++, Moderate; +++, Strong expression (Relative).

2.6 Figures: Chapter 2

Figure 2-1. Lmx1 phylogram.

Phylogram showing relatedness between LMX1A, LMX1B and their zebrafish orthologous proteins.

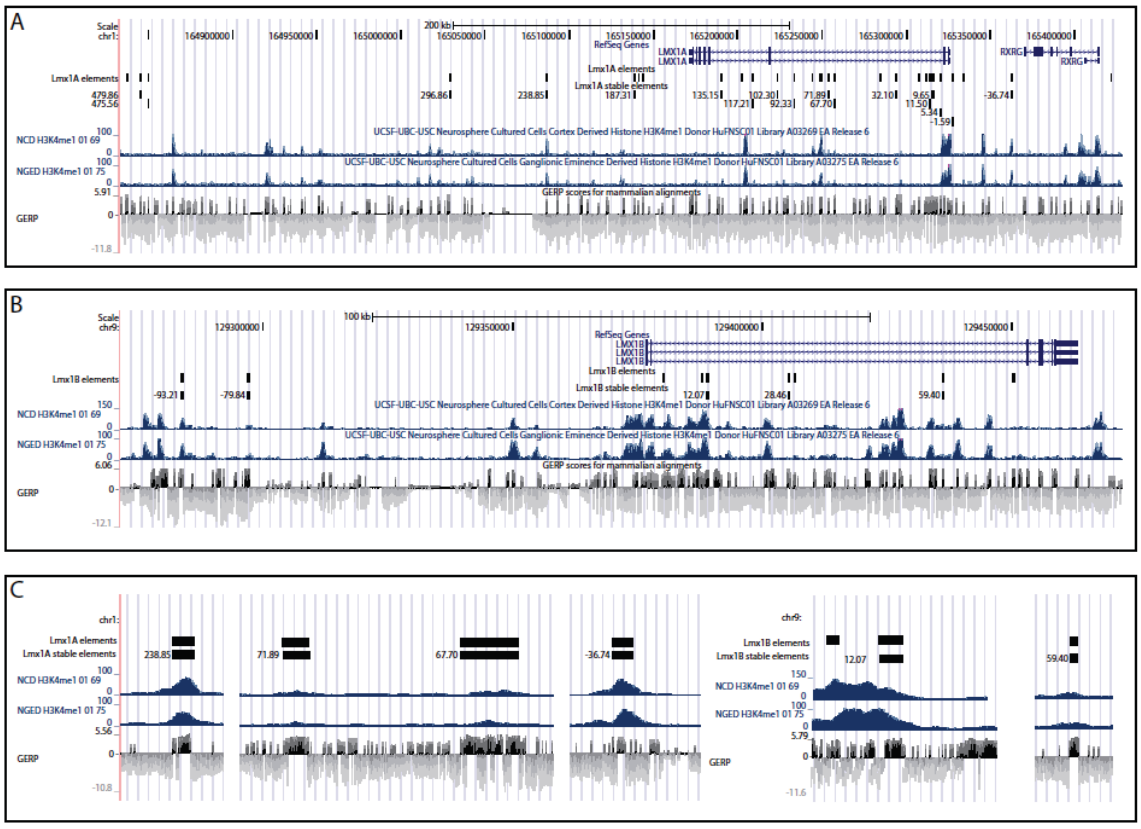
Figure 2-2. In situ hybridization showing the expression patterns of endogenous zebrafish LMX1A and LMX1B orthologs.



Expression of *lmx1a* (A-D), *lmx1a-like* (E-H), *lmx1ba* (I-L) and *lmx1bb* (M-P) are shown, assayed at 48 hpf and 72 hpf.

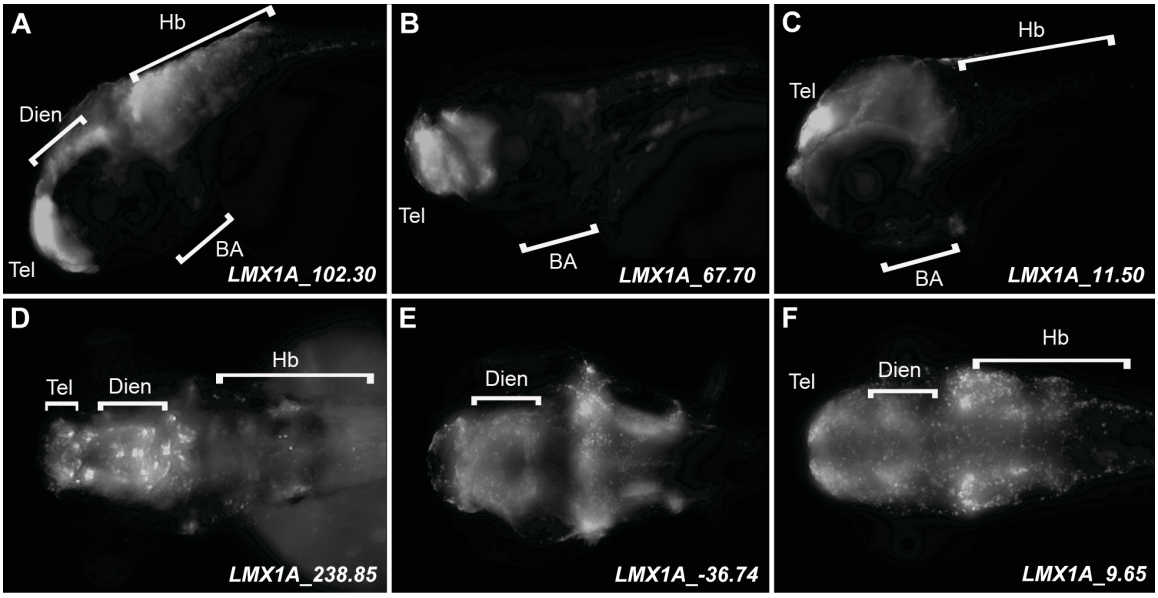
Abbreviations: Am – Amacrine neurons, CG – Cranial Ganglia, DD – Dorsal Diencephalon, dHB – Dorsal Hindbrain, Ep – Epiphysis, H – Heart, Hb – Hindbrain, Mb-Hb – midbrain-hindbrain boundary, OV – otic vesicle, RL – rhombic lip, RNa – anterior raphe nuclei, RNp – posterior raphe nuclei, VD – ventral diencephalon. Anterior is shown to the left in each image.

Figure 2-3. LMX1A and LMX1B loci and the noncoding conserved sequences selected for analysis.



LMX1A (A) and *LMX1B* (B) genomic loci displaying the selected sequences and their corresponding GERP sequence conservation tracks. H3K4me1 ChIP-seq signal from two types of cultured neurospheres (cortex derived and ganglionic eminence derived) are included to highlight the substantial overlap observed between conservation and high H3K4me1 signal intensity. C) Enlarged example intervals show the local sequence conservation within amplicons. The names of REs indicate the approximate distance in kb from the transcriptional start site of each gene.

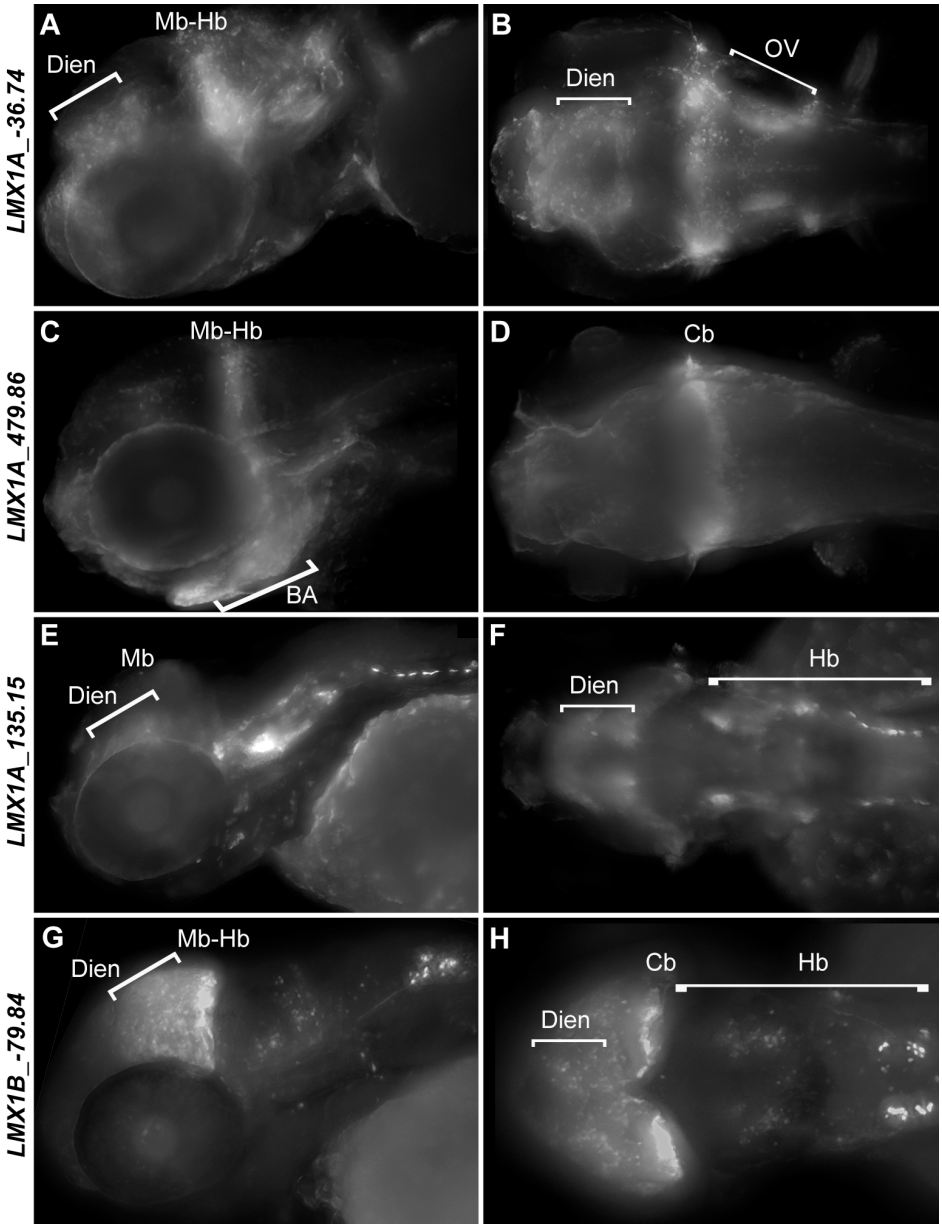
Figure 2-4. LMX1 enhancers drive reporter expression in the zebrafish diencephalon and telencephalon.



Representative images of LMX1 sequences that drive expression in the diencephalon and telencephalon. A-C, lateral images. D-F, dorsal images.

Abbreviations: BA – branchial arches, Hb – hindbrain, Dien – diencephalon, Hb – hindbrain, Tel – telencephalon.

Figure 2-5. Multiple LMX1 elements drive expression in the hindbrain and midbrain-hindbrain boundary.



Representative LMX1 transgenic zebrafish lines showing reporter expression in the hindbrain and midbrain-hindbrain boundary. Anterior is shown to the left in each panel. A, C, E, G are lateral images. B, D, F, H are dorsal images.

Abbreviations Cb – cerebellum, Dien – diencephalon, Hb – hindbrain, Mb – midbrain, Mb-Hb – midbrain-hindbrain boundary, OV – otic vesicle.

CHAPTER 3

DEVELOPMENT OF A PRIMITIVE CLASSIFIER TO DISCOVER A HINDBRAIN REGULATORY VOCABULARY

3.1 Introduction

In this chapter I will discuss our work towards extending our ability to identify enhancers from a locus specific approach to a tissue-specific approach. We asked whether regulatory signatures (vocabularies) could be uncovered from a complex cellular substrate, the central nervous system. In particular, we set out to determine the sequence basis of regulatory control in the hindbrain (Hb). The Hb, or rhombencephalon, is the most primitive part of the human brain, and likely evolved from a homologous structure present in Urbilateria around 550 million years ago (Ghysen, 2003). It includes the cerebellum, pons, and medulla oblongata, which are structures that control functions as fundamental and diverse as respiration, heart rate, reflex and voluntary movements. Impaired Hb development and function are associated with many disorders such as autism, ADHD (Attention Deficit Hyperactivity Disorder), schizophrenia, cerebral palsy, and various sleep disorders (Berquin et al., 1998; Aston-Jones, 2005; Andreasen and Pierson, 2008)].

As with other diseases and phenotypes, most variants identified by genome-wide association and sequencing population studies are found in non-coding regions of the genome, and therefore suspected to play a role in regulatory control (Cooper and Shendure, 2011). Understanding the gene regulatory landscape of the human genome in Hb development and structure is an important step towards uncovering the non-coding substrate of the genomic component of brain disorders.

We introduce a machine learning approach, based on the enrichment of transcription factor binding sites (TFBSs) in sequences, which are capable of accurately identifying enhancers which drive expression in the developing Hb. Our classifier performs very well in *de novo* discovery of Hb enhancers, with 88% (30/34) of computational predictions validated *in vivo* using transgenic zebrafish reporter assays. We also analyze the impact of small collections TFBSs on the Hb function of the host enhancers, and present a catalog of 40,000 putative Hb enhancers in the human genome. In summary, our data show how the application of effective computational methods for enhancer prediction can greatly improve our understanding of the gene regulatory networks controlling human development and disease.

3.2 Results

3.2.1 Building a training set of hindbrain enhancers

In order to construct a model for Hb enhancer activity, we first compiled a data set of 211 enhancers for which Hb activity has been validated *in vivo* with reporter assay systems in transgenic mice and zebrafish (Appendix 3). Most of these sequences (n = 192) were obtained from the VISTA Enhancer Browser (Visel et al., 2007) and an additional 20 enhancers were identified in our laboratory in the context of ongoing in vivo transgenic enhancer screens in zebrafish. This set of Hb experimentally proven enhancers bears genomic features consistent with other enhancer sets. The GC and repeat-content is close to the genome averages (Figure 3-1). Thirty-nine percent of the Hb enhancers in this catalog are intronic and 61% are intergenic, displaying a genomic distribution close to expected (for comparison, 44% of enhancers in the VISTA database are intronic). On the other hand, these Hb enhancers are especially well conserved among vertebrates – 99% of the Hb enhancers are conserved between human and mouse genomes, and 82% are also conserved between

human and chicken genomes. The average PhastCons evolutionary conservation score (Siepel et al., 2005) of Hb enhancers is 1.6, which is significantly higher than the corresponding scores of heart and limb enhancers (0.5 and 1.2, respectively; Wilcoxon rank-sum test P-value << 0.001). Enhancers driving expression in the nervous system frequently direct expression in one or more additional tissues or developmental stages. Accordingly, eighty percent of Hb VISTA enhancers also direct transcription in other tissues, such as midbrain (49%), forebrain (33%), neural tube (43%), and limb (8%), suggesting that the same elements may play pleiotropic roles in expression, and thus that regulatory lexicons may not always be discrete.

3.2.2 Designing an enhancer classifier

There is broad interest in determining the extent to which computational power can be used to elucidate how transcriptional regulatory instructions are encrypted in primary DNA sequence. The increased volume of genomic sequence-based data sets far exceeds our present capacity to impute biological value to primary sequence and variation therein, particularly in noncoding sequence. We previously developed a linear regression approach that relies on sequence patterns to accurately predict sequences with similar regulatory activity in the human genome de novo beginning with a small catalog of known heart enhancers (Narlikar et al., 2010). Since then, a similar method based on support vector machines (SVMs) and primitive short sequence segments (k-mers) has also performed well in classifying enhancers from different expression domains, including forebrain- and midbrain-derived ChIP-seq data sets (Lee et al., 2011). However, the SVM method was unable to accurately distinguish between different brain enhancer data sets. This was likely due in part to the cellular complexity of the tissues in which analyzed sequences chosen for the training

set drove expression. Therefore, although both the SVM and the linear regression method exhibited similar performances, we opted to combine the specificity of our original classifier with the advanced statistical model proposed by the latter approach (Lee et al., 2011). To this end, we constructed an SVM classifier which searches for overrepresented known TFBSs and de novo identified motifs, which we dubbed EnhSVM. We then used this strategy to determine if we could better discriminate among regulatory catalogs of CNS subdomains and extend this to define a classifier for the Hb, which currently has no ChIP-seq substrate available.

When applied to the collection of 11 tissue-specific experimentally validated sets of VISTA enhancers (forebrain, midbrain, hindbrain, neural tube, limb, heart, dorsal root ganglia, branchial arch, nose, cranial nerve, eye) our classifier was able to discriminate all enhancer sets from background genomic regions with accuracies exceeding 60% according to the area under the Receiver Operating Characteristic (ROC) curve (AUC) measurements in all cases (Figure 3-2). The majority of predictions produced by these models only overlapped predictions from related tissues, indicating that our method identifies cell type-restricted enhancer signatures. CNS enhancer classifiers (forebrain, midbrain, hindbrain, neural tube) performed better than the rest (Figure 3-2), and the Hb classifier displayed the highest AUC accuracy at 91%.

3.2.3 Refinement of a hindbrain classifier

The embryonic Hb forms along the anterior–posterior axis and is initially segmented into a series units called rhombomeres. The identity of these rhombomeres is correlated with domains of Hox gene expression and function, which in turn are determined by a gradient of retinoic acid along the anterior–posterior axis of the Hb (Schneider-Maunoury et al., 1998).

The most anterior rhombomeres contribute to the metencephalon, which in humans develops into the pons and cerebellum, while the most posterior rhombomeres form the myelencephalon, leading to the medulla oblongata in humans. In order to determine if we could further refine our classifier's predictive capacity, we separated the data set of Hb enhancers into 161 anterior and 153 posterior Hb enhancers based on expression patterns driven by the sequences in embryonic mice at developmental stage E11.5.

Although these sets of enhancers are highly overlapping with 80% of the sequences driving reporter expression in both domains, we hypothesized that simple functional clustering should result in increasingly homogeneous data sets, more suitable for our method. Other groups have shown that combinations of multiple classifiers trained on different subsets of a larger dataset often outperform single classifiers (Kittler et al., 1998). We predicted that this sort of functional clustering would improve our classifier. We trained and tested three independent Hb classifiers using a standard 10-fold cross validation setup on five random partitions of the data, using three slightly different data sets: the complete Hb data set, the subset of Hb enhancers that are active in the anterior Hb, and the subset of enhancers which functions in the posterior Hb. We found that no single classifier significantly outperforms the others. Instead, all three Hb classifiers achieved average AUCs of about 90%, with a true positive rate (TPR) of at least 47% at a false positive rate (FPR) of 5% (Figure 3-3 A).

3.3.4 Hindbrain enhancers harbor putative binding sites for transcriptional regulators of cell identity

Our Hb classifiers rely on sequence motifs representing TFBSs that facilitate distinction of Hb enhancers from random genomic sequences. We analyzed the

discriminatory power of individual motifs to reveal specific TFs likely to interact with Hb enhancers. All three Hb classifiers identified motifs that are known to bind the critical Hb TFs including MEIS1, NKX6-1, HOX family members, and POU protein family members among the 100 most relevant sequence features for identifying Hb enhancers (Waskiewicz et al., 2001; Nelson et al., 2005; Kiyota et al., 2008). Similarly, motifs known to bind SOX2, which is highly expressed in the Hb and has multiple roles in CNS development, were common to all three Hb classifiers (Appendix 4; Kelberman et al., 2008). Many of these motifs are specific to Hb development and function, and their relevance differs for analogous classifiers trained on data sets of enhancers specific to other tissues (compared, for example, with motifs relevant to limb and heart gene expression regulation, Appendix 4; Figure 3-3).

As expected, the distinct sets of Hb sequences, even if largely overlapping, showed slight differences in the contribution of each motif to the decision function of the corresponding classifier. For example, we observed differences in the relevance of the estrogen receptor ESR1 motif, which is specifically enriched among enhancers active in the posterior Hb. Thus, the motif for ESR1 is among the 100 most relevant sequence features for the Hb classifier focusing on posterior Hb, but not among the 100 most relevant sequence features for the other two Hb classifiers. Estrogen receptor-related proteins, which can bind ESR1-like motifs (Vanacker et al., 1999; Giguere, 2002), have previously been implicated in anterior–posterior brain segmentation (Bardet et al., 2005). The ability of the Hb classifiers to recover motifs corresponding to TFs with known functions in the Hb provides additional validation of our model. However, we must caution that it is unlikely that all computationally predicted motifs are bound by a TF. Moreover, even if they are, establishing the identity of the TFs binding to these motifs is not straightforward, since the binding affinity catalog of

TFs is not complete and many motifs are recognized by multiple TFs (Vavouri and Elgar, 2005).

In order to determine the specificity of the motifs with high discriminatory power in the Hb classifiers, we compared them with additional EnhSVM classifiers trained on published EP300 ChIP-seq datasets in forebrain, midbrain, and limb tissues (Visel et al., 2009). While a negligible fraction (<5%) of EP300 peaks is shared among all data sets, overlap among EP300 peaks for closely related tissues, such as forebrain and midbrain, was higher (15%–20%), consistent with a cell-type-restricted EP300 binding specificity. Less than 10% of the motifs are shared among the 50 most relevant sequence features between the ChIP-seq based classifiers. Additionally, less than 20% of these factors overlap with the motifs identified for Hb enhancers, highlighting the ability of our Hb classifier to specifically capture the Hb enhancer code. The fraction of TFBSs shared between the Hb and other brain classifiers included binding sites for MEIS1, NKX, SOX, and HOX homeobox domain factors, as well as ZHX2, a TF that is active in cortex development (Wu et al., 2009).

3.2.5 Genome-wide predictions identify novel hindbrain enhancers

Our training set is made up mainly of deeply conserved sequences (Visel et al., 2009), so to obtain a genome-wide map of putative human enhancers active in the Hb, we restricted our genome scan to sequences which are at least conserved among mammals (n = 337,000; Siepel et al., 2005). We completed independent searches using the anterior Hb (aHb), the posterior Hb (pHb), and the full Hb enhancer classifiers. Approximately 40% of the conserved sequences scored positively for at least one classifier (Figure 3-4), but only 12% (40,000) scored positively for all three (we dubbed the overlap set HbEns, as it represents the most reliable prediction of Hb enhancers). Seventy-seven of the HbEns (0.2%)

are known hindbrain enhancers from the VISTA Enhancer Browser (Visel et al., 2007), and 26,000 (60%) overlap ChIP-seq of known enhancer marks (H3K4me1 or H3K27ac). Likely a reflection of the similarity of the training data, we observed a large overlap among the highest scoring predictions obtained by each Hb classifier (Figure 3-5 B). Interestingly, the overlap correspondingly increases when raising the score cutoff, suggesting that sequence signatures for general Hb activity, rather than anterior or posterior Hb, dominate the decision function of all three classifiers.

The genomic distribution of the HbEns is similar to that observed for the training set. Approximately half of the candidate enhancers are intronic and half are intergenic (Appendix 2). HbEns are also fairly uniformly distributed with respect to the conserved sequences that served as the basis for the genome scans, with an average of four candidates per locus and a maximum of 102 in the case of PTPRD, a 2.3 Mb gene highly expressed in brain and recently associated with ADHD (Elia et al., 2010).

Compared with all scanned conserved sequences, HbEns are enriched within the loci of genes that are known to play a role in Hb development (P-value = 2.8 3 109, hypergeometric test). Moreover, higher scoring predictions are located significantly closer to genes associated with Hb development (Figure 3-5 C), indicating that our method preferentially identifies enhancers that are active in the Hb. Although all HbEns are, as defined by our search space, conserved among mammals, their level of evolutionary conservation is notably elevated. HbEns are significantly more conserved with respect to the conserved sequences that served as the basis for the genome scans (based on average PhastCons scores (Siepel et al., 2005), P-value $<2.2 \times 10^{-}16$, Wilcoxon rank-sum test). Additionally, 21% of HbEns are shared with chicken, 8% with frog, and 3% with zebrafish. Also, relative to the conserved sequences that served as the basis for the genome scans,

conservation in vertebrates is slightly, but significantly, enriched among HbEns (P-value 2.3 x 10^{-11} for the overlap with regions that are also conserved in chicken, Fisher's exact test). We also found a statistical enrichment of DNase I hypersensitive sites (HSS) identified in genomic DNA isolated from human fetal brain among HbEns (1.2-fold enrichment as compared with low scoring sequences, P-value < 2.2 x 10^{-16}, Fisher's exact test), while we do not observe any enrichment for DNase I HSS in other fetal tissues, such as heart and lung. Although, the hindbrain is only a subset of the complex tissue analyzed in fetal brain, and may refer to a different developmental stage, the enrichment in brain DNase I HSS corroborates our predictions as tissue-specific enhancers.

Finally, to evaluate the ability of our method to accurately define tissue-specific sequence patterns, we compared the distribution of predicted Hb enhancers with forebrain, midbrain, and limb enhancer predictions obtained in the same manner. In particular, we sought to verify that our predictions are not generally shared between different tissues, which would suggest a failed attempt to define a tissue-specific classifier. After we trained additional classifiers on the corresponding EP300 ChIP-seq enhancer sets, we found that there is <20% overlap between the top 5% of high scoring predictions (16% forebrain, 13% midbrain, 9% limb). This overlap is further reduced to 12% when comparing the top1% of high scoring predictions. This confirms our hypothesis that genome-wide predictions of classifiers trained on enhancers with different activities constitute largely disjoint sets, suggesting that the corresponding classifiers recognize sequence patterns involved in different biological functions.

3.2.6 The hindbrain classifier is a highly accurate predictor of hindbrain activity in zebrafish

In order to determine the accuracy of our method we set out to determine how frequently our predictions identify active Hb enhancers *in vivo*. We selected 55 sequences with a positive scaled summary Hb score for functional evaluation in a transgenic zebrafish reporter assay (Table 3-1). To avoid the introduction of biases based on genomic position, we included both intronic and intergenic sequences residing on 21 different human chromosomes (all except chr10 and Y). In addition, six low scoring sequences with a scaled summary Hb score less than zero were selected as likely ''negative'' predictions (Table 3-2). Predicted sequences may not identify all functional components within a complete enhancer, thus although our predictions were based on 100–200 bp sequence intervals, we designed primers to include ~200 bp flanking each side of the original sequence. The average size of all assayed amplicons was 485 bp (Appendix 2).

All sequences were tested for enhancer activity in the Hb using our established zebrafish transgenesis pipeline (Fisher et al., 2006 a and b; McGaughey et al., 2008). We define hindbrain expression as any expression in the CNS region that is posterior to the midbrain extending through the myelencephalon and delimited by the anterior portion of the spinal cord. Since the training set sequences directed expression in a number of non-Hb tissues, we do not require that expression is restricted to this region, and are therefore testing the sensitivity of the classifiers to Hb patterns rather than the specificity. The vast majority of constructs (51/55 putative Hb enhancers) directed reporter expression in some portion of the CNS in mosaic zebrafish embryos. Similarly, 6/6 low likelihood predicted sequences displayed mosaic signals in some part of the embryo, including portions overlapping the CNS.

All embryos that displayed reporter expression in mosaics were raised to maturity and crossed with AB zebrafish to determine which sequences could direct EGFP expression in the Hb. In total we identified two or more founders for 34 putative Hb enhancers, of which 30 (88%) founder sets displayed concordant expression in the Hb (Appendix 1), a predictive success rate that meets or exceeds prior rates of enhancer validation using both computational predictions as well as EP300 ChIP-seq-based predictions (Figure 3-6; Narlikar et al., 2010, 62%; Blow et al., 2010, 84%). In contrast, none of the six low likelihood controls displayed consistent expression in the Hb when passed though the germline (Appendix 1).

The patterns observed in stable lines displayed marked pleiotropy in their range of reporter expression both in Hb regions as well as in non-Hb regions, likely reflecting the heterogeneity of the training set. Figure 3-7 provides eight examples that illustrate the diverse patterns of expression observed in our validation set. Although the models trained on anterior and posterior sets of sequences did not appear to be particularly predictive of the relative position of Hb expression, we found that the resulting patterns could be grouped into categories displaying similar expression.

HB41, HB34, and HB02 share an expression profile that includes the cerebellum, part of the anterior Hb, in addition to varying levels of expression along the length of the Hb (Figure 3-7, A–C). However, HB02 also directs non-neuronal expression in the lens of the eye and myotome (Figure 3-7, C), which may be a result of position effects based on the site of amplicon insertion in the zebrafish genome as it was not observed in all stable lines. Some sequences, like HB15 (Figure 3-7, D), show expression in the Hb and very little extraneous expression. In contrast, HB25 and HB51 share a different expression profile displaying strong expression in the dorsal Hb as well as the tegmentum, a structure in the midbrain that is continuous with the medulla oblongata (Figure 3-7, E,F; Thisse and Thisse, 2004; Thisse et

al., 2004). HB10 shows distinct expression in the Hb, spinal cord, and dorsal diencephalon, as well as faint expression in the tegmentum and non-neuronal expression in the myotome and fin buds (Figure 3-7, G). In contrast to the distinct Hb expression seen in HB10, many domains within the CNS are faintly marked by reporter expression directed by HB50, including Hb neurons, cerebellum, tegmentum, dorsal diencephalon, and telencephalon (Figure 3-7, H).

The varied patterns of expression observed within the Hb validation set are consistent with the diverse nature of the motifs comprising the classifier. Addtionally, this result is expected given that the training set is comprised of sequences that displayed significant pleiotropy and included sequences that directed expression in an array of Hb subdomains, and non-Hb tissues. Consequently, we expected that TFBSs contained within these amplicons, and contributing to their prediction, would be diverse. However, we also anticipated that they would include sites for factors whose endogenous expression overlap with domains of reporter expression. *In vivo* validated Hb enhancer sequences are enriched for the 100 most relevant motifs for discriminating Hb enhancers compared with random sequences with similar GC content (Appendix 4). TFBSs for proteins in the POU, NKX, or PAX families, as well as LHX3 are especially common in our validation set (Table 3-2). Consistent with the *in silico* evaluations of TFBSs identified in HbEns collectively, factors in these families play critical roles in neuronal development. Furthermore, the observed reporter expression for each is largely consistent with previously published zebrafish expression patterns for one or more of the corresponding TFs.

POU domains are found in a large family of TFs and bind the consensus sequence ATGCAAAT (Verrijzer and Van der Vliet, 1993). They are expressed mainly in the CNS, and act as regulators of neurogenesis in zebrafish (Spaniol et al., 1996). Consistent with these

data, POU family TFBSs were the most commonly identified sites in our validation set and showed an enrichment of 2.6 over 600 randomly selected GC matched control sequences (P-value = 0.01, Fisher's exact test; Table 3-2) and many of our elements share expression domains with POU factors.

NKX proteins are necessary for the proper development of motor neurons in the hindbrain (Pattyn et al., 2003) and consistent with this role we see a significant enrichment (2.4; P-value = 0.005, Fisher's exact test; Table 3-2) for NKX family TFBSs in our validated set of Hb enhancers. Similarly, the PAX gene family comprises a large group of highly conserved TFs required for neuronal development (Wang et al., 2010; Thompson and Ziman, 2011), and 10/30 validated predictions contained at least one PAX family motif (enrichment of 1.8; P-value = 0.05, Fisher's exact test; Table 3-2). Finally, a number of sequences contain an LHX3 motif that binds a LIM domain TF with a role in neuronal specification (Cepeda-Nieto et al., 2005; Gadd et al., 2011), resulting in an enrichment of 4.6 (P-value = 0.0002, Fisher's exact test; Table 3-2). HB25 and HB51 both contain an LHX3 TFBS and share many overlapping domains of reporter expression, including in the Hb and spinal column, which is consistent with endogenous lhx3 expression (Figure 3-6 E,F; Appendix 1; Thisse and Thisse, 2004; Thisse et al., 2004). In contrast to our HbEns predictions, only one of the low likelihood controls contained any of these motifs, supporting their high predictive power in our model. Collectively, these data provide compelling evidence that the HbEns sequences may be important in regulating transcription in the developing Hb.

3.2.7 Tissue-restricted enhancer activity is due to the presence of specific transcription factor motifs

Our data suggest that TFBSs contributing to the Hb classifier might independently or collectively explain aspects of the observed regulatory control of the sequences within which they reside. We selected two sequences displaying Hb regulatory control (HB01 and HB16) to examine more closely, surveying the distribution of TFBSs within each predicted sequence. We then identified smaller sequence fragments for analysis in zebrafish based on the clustering of TFBSs therein. The full-length HB01 sequence directed distinct expression in the rhombomeres, as well as the midbrain Hb boundary, cranial ganglia, and dorsal diencephalon (Figure 3-8, B and E).

We amplified two smaller fragments (HB01_I and HB01_II) from within the full-length sequence based on the pattern of TFBSs clusters. HB01_I is a 56-bp sequence containing motifs for PITX2, CDX, CEBPG, NKX3-1, and BCL6 (Figure 3-8, A). Upon passage through the germline, HB01_I displayed broad reporter expression in the CNS (Figure 3-8, C and F). This pattern encompasses the expression domains also marked by the full-length HB01. The expanded expression domains of HB01_I could reflect the increased efficiency of TFBSs being placed closer to the minimal promoter (Nolis et al., 2009). It may also reflect the absence of other regulatory sequence motifs within or beyond the initial predicted interval which otherwise act in the full-length construct to mediate transcriptional activity (Gompel et al., 2005). Notably HB01_II, which is 93 bp and contains motifs for HNF3, POU2F1, NKX2-5, MYOG, SOX10, and HMGA1 (Figure 3-8, A) did not show any mosaic expression, and was determined to be insufficient for enhancer activity in the Hb in this assay (Figure 3-8, D and G).

Similarly, HB16 displays prominent expression in the dorsal Hb and fainter expression in the ventral Hb and lateral tegmentum (Figure 3-8, I and M). Once again we amplified three short fragments from within the initially predicted sequence based on TFBS clustering (Figure 3-8, H). HB16_I is a 29-bp fragment containing a GATA1 motif; HB16_II is 54 bp in length and contains MEOX2, NKX6-1, EN1, TAL1, NKX2 family, JUN, and PAX4 motifs; and HB16_III is a 23-bp fragment encompassing a HOXA4 motif (Figure 3-8, H). Upon passage through the germline, both HB16_I and HB16_II directed expression in the Hb (Figure 3-8, J,K,N,O). In contrast, HB16_III only drove expression in the myotome of stable lines (Figure 3-8, L and P). Notably, the reporter expression in the Hb neurons and the lateral tegmentum directed by HB16_I are similar to those of the endogenous gata3 (Figure 3-8, J and N; Thisse and Thisse, 2004; Thisse et al., 2004). This pattern is consistent with expression directed by full-length HB16.

Furthermore, HB16_II directs expression along the entire length of the ventral and medial Hb and spinal column (Figure 3-8, K and O). As such, it overlaps much of the Hb domain marked by HB16 and resembles the endogenous expression of nkx6 family, nkx2 family, and tal1 RNA (Thisse and Thisse ,2004; Thisse et al., 2004; Binot et al., 2010). The observed reporter expression in the tegmentum is also consistent with endogenous tal1 RNA (Thisse and Thisse, 2004; Thisse et al., 2004). A potential role for the JUN TFBS identified in HB16 is not obvious but these factors display broad expression throughout the CNS and may account, at least in part, for expression domains extending dorsally. Although not conclusive, these data suggest that the expression of TFs corresponding to motifs contributing to our classifier are consistent with their predicted biological roles in modulating expression in the Hb and show that enhancers can be further broken down into their

component TFBS fragments while continuing to faithfully drive reporter expression in the predicted tissue.

3.3 Conclusions

As we have previously discussed, the exquisite orchestration of transcriptional control is essential for the normal development and homeostasis of multicellular organisms, however systematic identification of sequences responsible for these activities has proven a significant challenge. Although sequence constraint has been used with significant success, it can impute little regarding the likely biological activity of any identified sequence. Similarly ChIP-seq profiling of TFs, histone modifications, and transcriptional co-activators such as EP300 has recently emerged as a powerful tool for the identification of enhancers active in various tissues; however, not all enhancers are captured by affinity-based methods, and not all cell types are amenable to these assays. Recent efforts to identify overrepresented sequence motifs have proven increasingly powerful, allowing the elucidation of early language structure for regulatory control in specific tissues (Narlikar et al., 2010; Lee et al., 2011).

We have integrated these computational strategies, employing machine learning to train a sequence-based classifier on a set of largely published *in vivo* validated enhancers in the Hb. The result is a highly accurate predictor of enhancer activity in the Hb. When applied to the human genome, 88% (30/34) of sequences demonstrate Hb regulatory control when assayed *in vivo*. In contrast, in prior studies that identified sequences as being deeply conserved only ~8% were observed to drive expression in the Hb (Pennacchio et al., 2006). The motifs identified by our classifier frequently represent TFBSs for factors with known roles in regulating transcription in the Hb and with endogenous expression patterns

overlapping with that of reporter expression. Furthermore, we show that, consistent with our classifier, smaller sequences containing clusters of TFBSs (~30 – 100 bp) contributing to predicted Hb regulatory control can account for aspects of Hb regulatory expression observed in the original (~500 bp) sequence from which they were derived.

Although the vocabulary described is an effective predictor of Hb activity, we observed pleiotropy among Hb domains marked by reporter expression as well as expression in domains outside the Hb, including non-neuronal tissues. These observations are consistent with the complexity of vertebrate enhancers known to display a broad expression pattern across multiple tissues (Visel et al., 2007). It is particularly important to keep in mind that the Hb enhancers in our training data set were not exclusively expressed in the Hb, but largely displayed multi-tissue expression patterns. From the sequence analysis perspective, our training set contained a large group of Hb enhancers and several smaller clusters of other expression subdomains. The non-Hb signatures in our training set likely created a plethora of misleading signals confusing the classifier. However, the high Hb validation rate of HbEns reflects the ability of the classifier to sensitively extract the Hb sequence encryption from the noisy input data set. Knowing that Hb sequence encryption often resides within enhancers with broad expression patterns and does not represent a code of exclusive Hb expression, we expected the observed pleotropic expression of experimentally assayed HbEns sequences.

As additional support for the utility of our model, we find that our predicted Hb enhancers are enriched for a particularly large number of CNS TFBSs compared with TFs known to be active in other tissues. Our experimental data also suggest that Hb enhancers can be divided into independent functional subunits with similar activities but different sequence structures—an observation that highlights flexibility of the Hb sequence encryption with potential for adaptation to additional functions and the use of different activation

mechanisms. The observed biological behaviors of these TFBS clusters were consistent with the known patterns of expression of TF family members predicted to bind them. This raises the possibility that retraining algorithms using subsets of training or predicted sequence sets may define the sequence grammar that is specific to individual Hb sub-domains and cell types.

Computational methods are becoming increasingly powerful tools for enhancer prediction. Here we have shown that experimental validation rates for computer-learning algorithms are comparable to those achieved by experimental ChIP-seq predictions. Computational predictions can be similarly independently correlated with the presence of features known to be present in active enhancers such as known TFBS motifs, specific histone marks, and increased conservation.

This study demonstrates that, in addition to the sequence substrate provided by genome-wide ChIP-based strategies, the published literature may serve as a valuable entry point for such analyses of regulatory elements. We demonstrate that even a relatively small curated experimental data set can provide significant insight into the regulatory lexicon of a highly complex anatomical structure like the Hb, and that this vocabulary can likely be dissected and improved in subsequent cycles of investigation and/or by the refinement of the substrate on which it is trained. Therefore, this study adds to the ongoing project of genome annotation by identifying sequences that have a functional role in the Hb. The development of regulatory language is a pivotal step in the prediction of functional variation by inspection of the primary sequence and as such this study makes a significant first step in the development of a hindbrain lexicon.

3.4 Methods

3.4.1 Tissue-specific enhancer models

We extracted 771 human sequences from the VISTA Enhancer Browser (Visel et al., 2007) with validated in vivo enhancer activity in 23 tissues. We were able to retrieve at least 29 sequences each for 11 of these tissues. Each enhancer model was trained to distinguish between enhancers specific for a given tissue and other noncoding sequences, randomly drawn from the noncoding human sequence, with length, GC, and repeat-content distributions similar as those observed for the enhancers. The decision of the corresponding classifier was based on the presence or absence of two different types of motifs: 775 corresponding to binding specificities of vertebrate TFs compiled in public databases (TRANSFAC and JASPAR; Matys et al., 2003; Bryne et al., 2008), and 20 short sequence patterns enriched among the set of enhancers, identified with PRIORITY (Narlikar et al., 2007), which should account for the binding of unknown TFs or TFs with unknown binding specificities. Thus, each sequence was represented as a feature vector indicating the number of matches per base pair to each of these motifs, computed using MAST (Bailey and Gribskov, 1998). We built the classifier using linear SVMs (implemented in libsvm; Chang and Lin, 2011), assuming no prior knowledge of TFs active in the different tissues, with the goal being to discover them using the feature weights learned by the classifier.

3.4.2 Hindbrain enhancer models for mouse, chicken, frog, and zebrafish

Orthologous regions of the human Hb enhancer training set were identified using the liftOver utility from the UCSC Genome Browser (Karolchik et al., 2008). We discarded mapped sequences longer than 5 kb. We successfully mapped 100%, 86%, 74%, and 47% of the 211 Hb enhancers onto the mouse (mm9), chicken (galGal3), frog (xenTro2), and

zebrafish (danRer5) genomes respectively. For each enhancer in the training set, 10 controls with similar length, GC, and repeat-content were randomly drawn from the noncoding portion of the corresponding genome.

3.4.3 Forebrain, midbrain, and limb enhancers identified using ChIP-seq

Genomic regions enriched for EP300 binding in mouse forebrain, midbrain, and limb tissues were extracted from Supplemental Tables 2–4 of Blow et al. (2010). We identified orthologous regions of the mouse coordinates with the liftOver utility from the UCSC Genome Browser (Karolchik et al., 2008). Sequences longer than 1 kb were discarded, resulting in a total of 2199 forebrain sequences, 1909 midbrain sequences, and 3155 limb sequences.

3.4.4 TFBS mapping, association and enrichment

Putative TFBSs were identified de novo by searching the predicted enhancer sequences with MAST (Bailey and Elkan, 1994) for 775 motifs in the TRANSFAC Release 2009.2 (Matys et al., 2006) and JASPAR (Bryne et al., 2008) databases. MAST was run independently on each individual sequence with default setup and parameters. The identity of the TFs binding to de novo motifs was queried using STAMP (Mahony and Benos, 2007) and JASPAR (Bryne et al., 2008). TF annotation for known TFBSs was obtained from TRANSFAC, JASPAR, and the Broad Institute MsigDB (Subramanian et al., 2005). Overrepresented TFBSs were determined by comparing the occurrence of the motifs among query sequences and background genomic sequence, and applying Fisher's exact test using a P-value threshold of 0.05. When indicated, we adjusted the P-values for multiple testing using the procedure suggested by Benjamini and Hochberg (1995).

3.4.5 Extracting homogeneous Hb enhancer data sets

Hb enhancers tend to drive expression in multiple tissues, and even show heterogeneous patterns of expression within the Hb. As a result it is unlikely that we would be able to identify a unique set of sequence features representing all Hb enhancers. Thus, similar to the approach taken in Narlikar et al. (2010), we selected a large subset of these sequences sharing homogeneous sequence features as an attempt to reduce the sequence heterogeneity among the 212 human Hb enhancers. For this purpose, we repeated the 10-fold cross-validation on five random partitions of the Hb enhancer data set as well as on that of the corresponding controls, and selected only those Hb enhancers that were classified as such in at least 50% of the times in which they were tested for the final training set. Therefore, the final human Hb enhancer data set contained 124 sequences. The performance of the classifiers was evaluated in a 10-fold cross validation, using the area under the ROC curve (AUC). AUC values range from 0.5 (random discrimination) to a theoretical maximum of 1.

3.4.6 Creating linear SVMs

Training a linear SVM classifier is equivalent to solving the following constrained optimization problem (Shawe-Taylor and Cristianini, 2002):

Given the training samples $T = \{(x_i, y_i) | x_i \in \mathbb{R}^p, y_i \in \{-1,1\}\}_{i=1}^{n}$ find the values of w, b and ξ_i that minimize $\frac{1}{2} w^T \cdot w + C \sum_{i=1}^{n} \xi_i$

satisfying the constraints $y_i(w^T \cdot x_i + b) \geq 1 - \xi_i \; \forall i = 1, \ldots, n$

and $\xi_i \geq 0 \; \forall i = 1, \ldots, n$

The decision function of the classifier for an unknown sample x is given by

$$f(x) = \text{sign}(w^T \cdot x + b)$$

The dual form of this problem is: Given the training samples

$$T = \{(x_i, y_i) | x_i \in \mathbb{R}^p, y_i \in \{-1,1\}\}_{i=1}^{n}$$

find the values $\{\alpha_i\}_{i=1}^{n}$ that maximize $\sum_i \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j$

satisfying the constraints $0 \leq \alpha_i \leq C \; \forall i = 1, \ldots, n$

and $\sum_{i=1}^{n} a_i y_i = 0$.

Samples x_i for which $a_i \geq 0$ are called support vectors.

The vector w can be computed in terms of α_i as $w = \sum_{i=1}^{n} a_i y_i x_i$

and, therefore, contains the weighted features of the support vectors.

3.4.7 SVM parameter selection

Linear SVMs have only one parameter, C, which controls the trade-off between errors on the training data and margin maximization. We found that the performance of the Hb enhancer classifier was relatively stable with respect to changes in C. We estimated C based on the training data as $\left[\frac{1}{n}\sum_{i=1}^{n}|x_i|\right]^{-2}$. Additionally, because the training data are unbalanced (there are 10 controls for each enhancer sequence), misclassifications are penalized differently depending on the class of sequences (controls and enhancers), proportionally to the total number of sequences in each class.

3.4.8 Motif rankings

After obtaining a linear SVM model, the weight vector w can be used to decide the relevance of each feature (Guyon et al. 2002). The larger $|w_j|$, the more important role of feature j in the decision function. We rank features, in our case, motifs, according to $|w_j|$. We

exclude de novo motifs from these ranks unless stated otherwise. It is important to note that this interpretation for w is only valid for linear SVMs.

3.4.9 Hindbrain genes

We identified a set of 787 human genes likely to be involved in Hb function by retrieving genes with relevant phenotypes from the Online Mendelian Inheritance in Man (OMIM) database (Amberger et al., 2009) and the corresponding orthologs of genes with pertinent annotation in the Mammalian Phenotype (MP) Browser at the Mouse Genome Informatics website, The Jackson Laboratory, Bar Harbor, Maine (http://www.informatics.jax.org).

3.4.10 Genome scans

We applied three human Hb enhancer models trained on (1) the complete Hb data set, (2) the subset of Hb enhancers that are active in the anterior Hb, and (3) the subset of enhancers which functions in the posterior Hb to scan sequences highly conserved across mammals using the Most Conserved Elements database from the UCSC Table Browser (Siepel et al., 2005). Noncoding conserved sequences were determined based on annotation in UCSC Known and RefSeq (Hsu et al., 2006; Pruitt et al., 2009). Sequences within 100 bp of each other were clustered together and clusters shorter than 100 bp were excluded from the analysis. Using classifiers trained on the orthologous sequences of the complete data set of human Hb enhancers, we utilized an analogous procedure to predict Hb-specific enhancers in the mouse, chicken, frog, and zebrafish genomes.

3.4.11 Scaled summary Hb score

Each scanned sequence is given three scores, anterior Hb score, posterior Hb score, and general Hb score; by the classifiers trained on the subset of Hb enhancers that are active in the anterior Hb, the subset of enhancers which functions in the posterior Hb, and the complete Hb data set, respectively. The scores are distributed in the range [-17,15], [-20,15], and [-22,15], respectively (Figure 3-4). Scores greater than zero correspond to putative enhancers active in the anterior Hb, in the posterior Hb, and in the (general) Hb, respectively. Approximately 130,000 sequences scored greater than zero for at least one of the classifiers, while 40,000 sequences scored greater than zero for all three. Scores for all classifiers are subsequently linearly scaled according to

$$score^* = \begin{cases} -\left(1 - \frac{score - score_{min}}{-score_{min}}\right), & if\ score < 0 \\ \frac{score}{score_{max}}, & if\ score \geq 0 \end{cases},$$

where $score_{min}$ and $score_{max}$ are the minimum and maximum scores obtained in the genome-wide scan, respectively. Finally, we define the scaled summary Hb score as the maximum between $score^*_{anterior\_Hb}$, $score^*_{posterior\_Hb}$, and $score^*_{general\_Hb}$.

3.4.12 Association between enhancer predictions and loci

In order to associate putative enhancers with genetic loci we defined gene loci in the human genome using known Gene and RefSeq annotation tracks available at the UCSC Genome Browser (November 2011). One or more overlapping transcripts, prohibiting overlap among different loci, defined each locus. Putative Hb enhancers were associated with loci based on genomic proximity. Therefore, our interpretation assumes that each putative Hb enhancer targets the nearest genetic locus.

3.4.13 DNase I hypersensitivity and histone modification comparisons

We compared our putative Hb enhancers with human fetal brain, heart, and lung DNase I hypersensitive peaks from http://nihroadmap. nih.gov/epigenomics/. H3K4me1 and H3K27ac peaks were obtained from http://genome.ucsc.edu/ENCODE/ (The ENCODE Project Consortium 2011) and correspond to ChIP-seq from all human cell lines available at the time of study.

3.4.14 In vivo validation

Candidate Hb enhancers for validation were selected randomly from positively scoring sequences with rank less than or equal to ~40,000. Controls were selected among sequences that scored among the bottom 1% (i.e., rank greater than or equal to ~334,000) for all classifiers. Zebrafish were maintained as previously described (Kimmel et al., 1995; Westerfield, 2000). Predicted enhancers were amplified by PCR from human genomic DNA and cloned using Gateway Technology (Invitrogen). PCR fragments were TA-cloned into the pCR8/GW/TOPO vector (Invitrogen) then TOPO-cloned using attL1 and attL2 sites into the pT2cfosGWvector for injection into zebrafish embryos. Short fragment sequences for HB01 and HB16 were synthesized as double-stranded oligos, A overhangs added, then cloned as predicted enhancers. At least 100 embryos were injected per construct at the two-cell stage with tol2 transposase as previously described (Fisher et al., 2006 a and b).

Injected embryos were screened for GFP expression in the CNS at 24 and 48 hpf. Those showing CNS expression were raised to adulthood and crossed to AB zebrafish. G1 embryos were screened for Hb expression at 24, 48, and 72 hpf. GFP positive embryos were live imaged at 72 hpf using a Carl Zeiss Lumar V12 Stereo microscope with AxioVision version 4.8 software. Embryos were fixed in 4% PFA (Sigma) overnight then post-fixed in

100% acetone (JT Baker) and washed in PBS with 0.5% Tween. Embryos were blocked in 10% goat serum and 1% BSA for two hours, then incubated with chicken anti-GFP (Invitrogen A10262, 1:1000) overnight. After washing, Alexa Fluor 488 goat anti-chicken IgG (Invitrogen A11039, 1:3000) was added and incubated overnight. After washing, embryos were stored in 80% glycerol at 4°C for future imaging.

3.5 Tables: Chapter 3

Table 3-1. Characteristics of Hb enhancers tested in vivo.

Name	Coordinates (hg18)	Size (bp)	Scaled summary Hb score	Overall rank
HB01	chr14:47542553-47542700	147	0.37	3221
HB02	chr10:11095875-11096151	276	0.24	12498
HB03	chr14:46003441-46003574	133	0.20	19431
HB04	chr12:85151257-85151367	110	0.28	8718
HB05	chr7:13451899-13452047	148	0.42	1726
HB06	chr2:206786161-206786276	115	0.51	670
HB07	chr10:105298975-105299102	127	0.63	188
HB08	chr5:88758665-88758837	172	0.30	6890
HB09	chr4:66889195-66889310	115	0.60	274
HB10	chr13:53651033-53651240	207	0.33	5048
HB11	chr14:27141446-27141697	251	0.25	11771
HB12	chrX:114374399-114374510	111	0.25	11475
HB13	chr1:206309342-206309511	169	0.27	9374
HB14	chr13:35635698-35635867	169	0.17	25561
HB15	chr16:13015785-13015896	111	0.31	5811
HB16	chr14:39591833-39591950	117	0.25	12128
HB17	chr6:2327522-2327644	122	0.24	13177
HB18	chr2:79807546-79807739	193	0.37	3091
HB19	chr15:62881358-62881475	117	0.33	4886
HB20	chr15:86535910-86536025	115	0.30	6480
HB21	chr12:6751734-6751913	179	0.32	5543
HB22	chr20:43405575-43405696	121	0.29	7400
HB23	chr5:139417800-139417932	132	0.22	16066
HB24	chr20:39964905-39965010	105	0.23	14041
HB25	chr2:117335633-117335763	130	0.27	8799
HB26	chr3:178842751-178842854	103	0.25	11706
HB27	chr18:48175265-48175412	147	0.15	32683
HB28	chr3:32423887-32423998	111	0.16	29741
HB29	chr17:35808909-35809107	198	0.15	32747
HB30	chr14:56583170-56583278	108	0.17	26867
HB31	chr13:52148456-52148606	150	0.15	33727
HB32	chr5:151560335-151560535	200	0.17	26389
HB33	chr16:84681714-84681816	102	0.15	32214
HB34	chr7:153188989-153189115	126	0.28	8604
HB35	chr11:102496520-102496626	106	0.23	13715
HB36	chr10:73516897-73517093	196	0.31	6018

HB37	chr6:49234872-49235047	175	0.25	11954
HB38	chr14:30318079-30318180	101	0.22	15269
HB39	chr22:36334856-36334957	101	0.61	233
HB40	chr2:232283281-232283418	137	0.25	11599
HB41	chr5:67641868-67641990	122	0.33	4882
HB42	chr10:117345190-117345362	172	0.20	18882
HB43	chr12:90547556-90547657	101	0.32	5241
HB44	chr22:34348464-34348585	121	0.17	26265
HB45	chr15:66897239-66897352	113	0.29	7175
HB46	chr16:49588251-49588360	109	0.20	20054
HB47	chr1:104891296-104891470	174	0.20	19179
HB48	chr3:142944914-142945039	125	0.21	18453
HB49	chr9:137002680-137002785	105	0.33	5021
HB50	chr17:32695070-32695176	106	0.19	22380
HB51	chr4:21785966-21786074	108	0.18	24045
HB52	chr2:30107189-30107302	113	0.16	29444
HB53	chr19:34550750-34550865	115	0.16	30856
HB54	chr15:83868753-83868853	100	0.15	34602
HB55	chr3:82099907-82100045	138	0.17	25603
HBN01	chr2:213474999-213475102	103	-0.82	336738
HBN02	chr4:16218968-16219081	113	-0.84	336739
HBN03	chr11:62420195-62420307	112	-0.69	336726
HBN04	chr16:1429021-1429130	109	-0.71	336730
HBN05	chr3:129750976-129751160	184	-0.55	336659
HBN06	chr9:109890045-109890169	124	-0.57	336675

Coordinates and scores of putative Hb enhancers and controls tested *in vivo*. Element name, prediction coordinates, size of prediction, scaled summary Hb score, general rank among all scanned sequences, and rank with respect to the HbEns set.

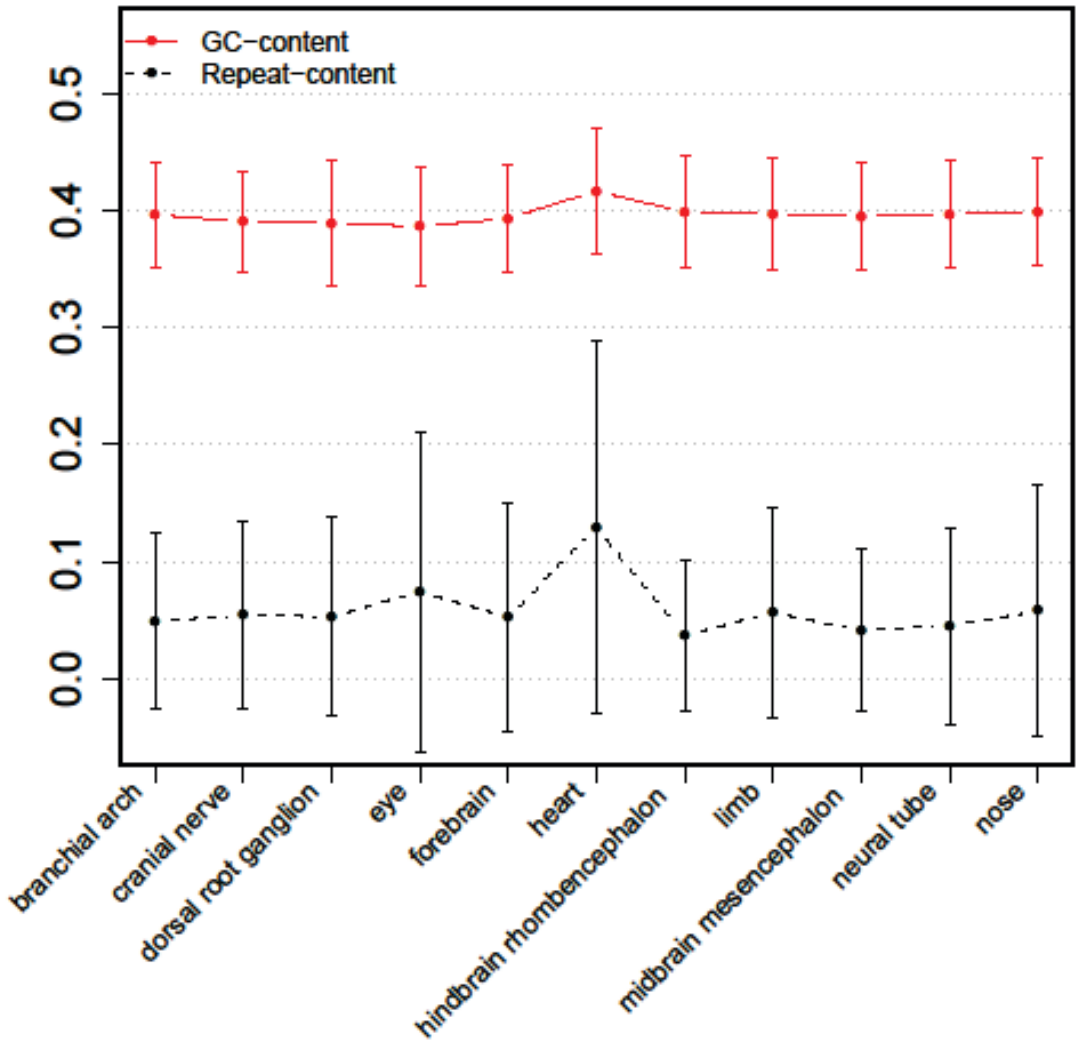
Table 3-2. Enrichment of motifs identified by Hb classifiers.

TF family	Validated Hb Enhancers	Random Controls	Relative enrichment	P-Value
POU	14/30	153/600	2.6	0.01
NKX	9/30	66/600	2.4	0.005
PAX	10/30	114/600	1.8	0.05
LHX	9/30	39/600	4.6	0.0002

Enrichment for the five most common TFs and TF families located in validated enhancer sequences. 20 GC matched random controls were used for each enhancer sequence. P-value is computed using Fisher's exact test.

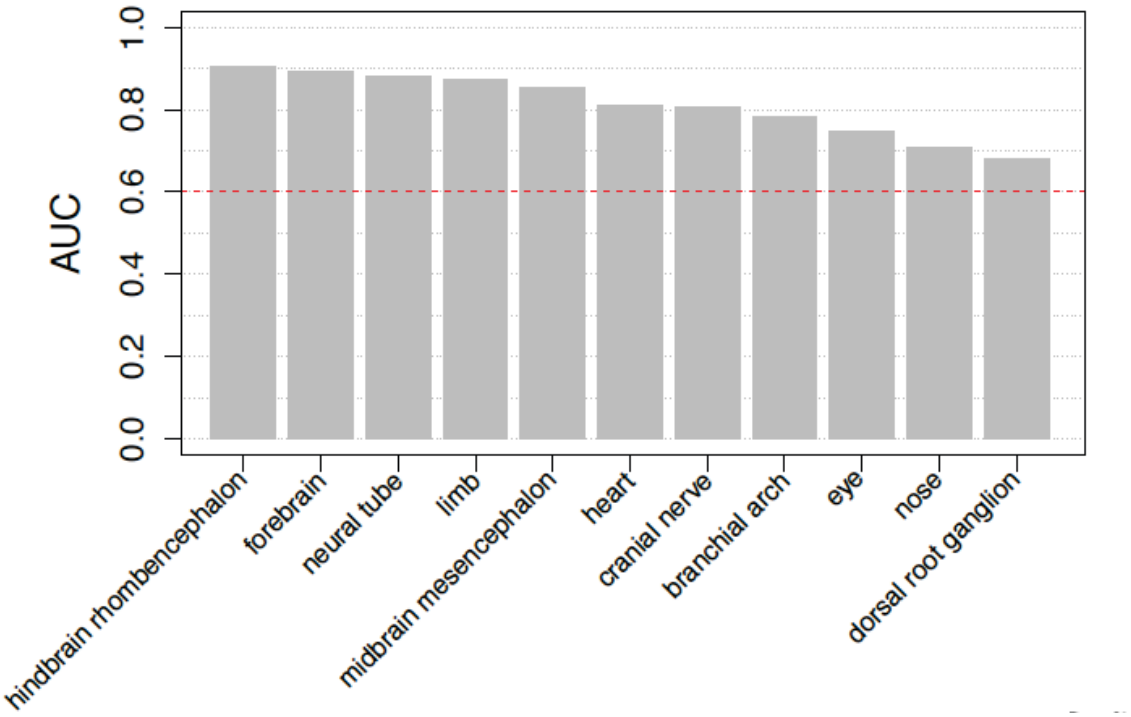
3.6 Figures: Chapter 3

Figure 3-1. GC and repeat content of Hb enhancers do not differ from enhancers driving expression in other tissues.



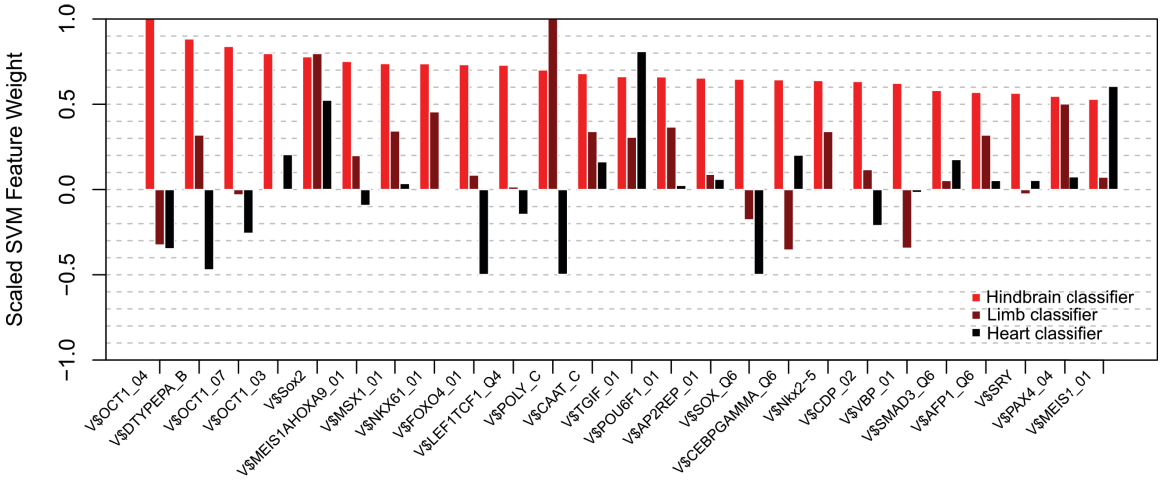
Analysis of GC content (red) and repeat content (black) proportion of different tissue-specific enhancers datasets. Proportions are approximately the same across all tissues examined.

Figure 3-2. Hindbrain classifier performs well relative to other classifiers in distinguishing tissue-specific enhancers from background genomic sequence.



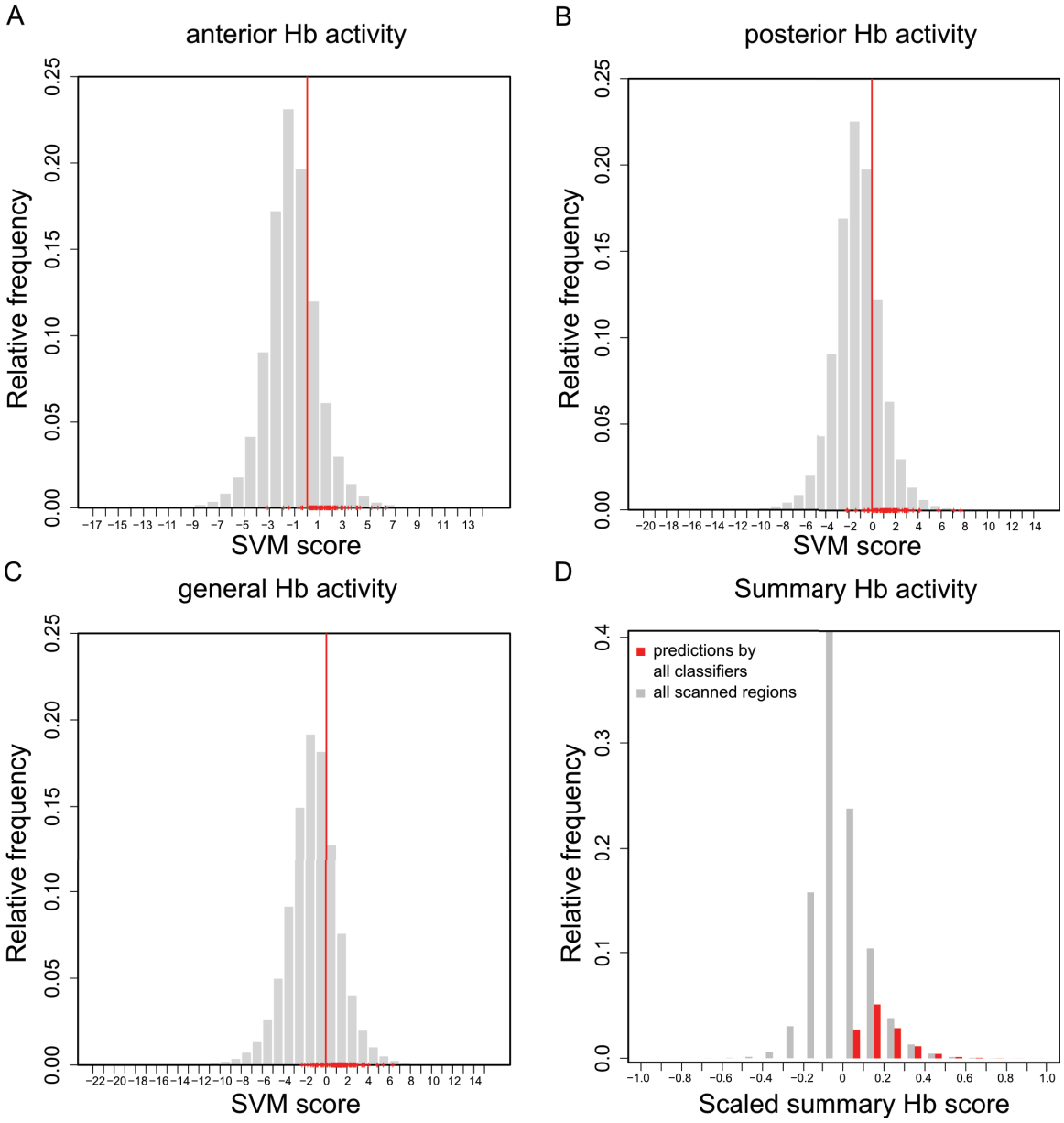
Area under the Receiver Operating Characteristic (ROC) curve (AUC) for 11 tissue-specific enhancer classifiers, each trained on at least 29 sequences extracted from the VISTA Enhancer Browser. All classifiers yielded reasonable performances (AUC $\geq$ 0.6, red dotted line). The highest AUC values are achieved for CNS tissues, including Hb.

Figure 3-3. Weights of the top motifs for hindbrain classifier relative to weights in limb and heart.



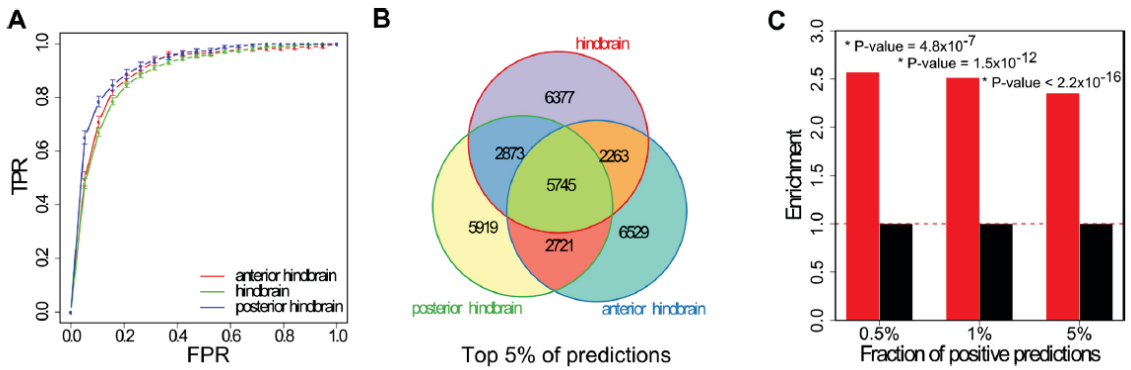
Scaled weights for the 25 most relevant motifs for the human Hb classifier trained on the complete Hb dataset (red). For comparison, we include the scaled weights for the same motifs, as determined by two additional classifiers trained on limb (dark red) and heart (black) human enhancers from the VISTA enhancer browser. The names of the motifs are listed near the baseline of the graph. Weights w_j have been linearly scaled to the intervals [0,1], if $w_j \geq 0$, and [-1, 0], if $w_j < 0$, according to $\frac{w_j}{w_{max}}$ and $-\left(1 - \frac{w_j - w_{min}}{-w_{min}}\right)$, respectively, and assuming that the range of the scores is $[-\infty, \infty]$. Motifs whose presence is relevant for identifying putative enhancers will have scaled weights close to 1, while motifs whose absence is relevant for identifying putative enhancers will have scaled weights close to -1.

Figure 3-4. Distribution of SVM scores obtained from each Hb classifier.



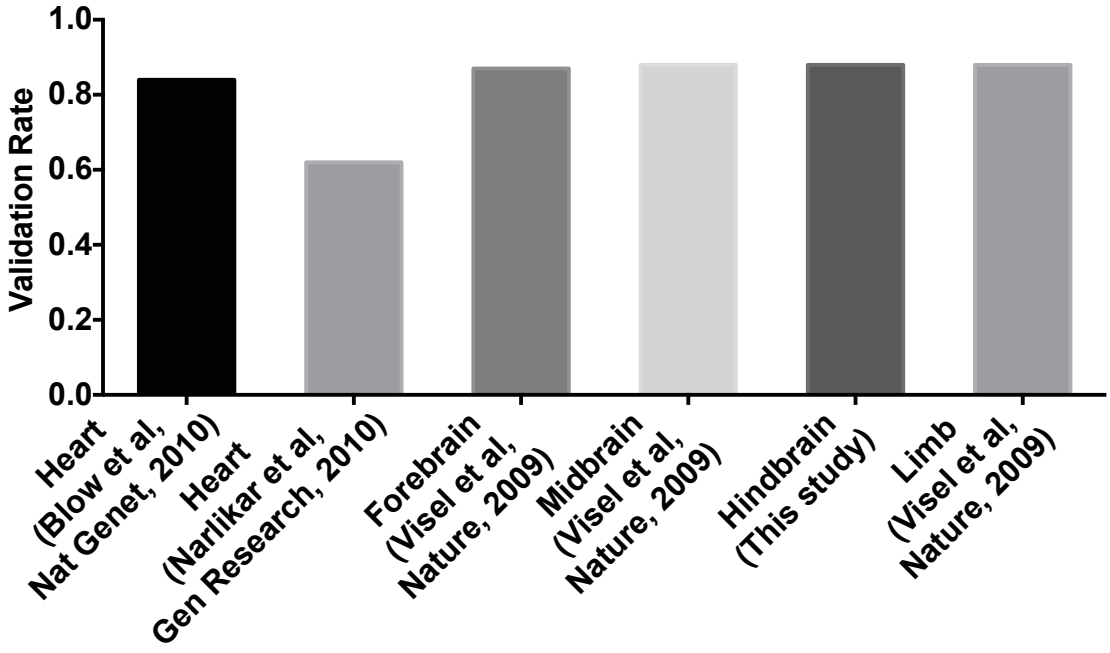
A, B, C) Distribution of scores for the genome-wide scans by the three human Hb classifiers, trained with overlapping subsets of the entire set of human Hb enhancers. A score of zero is used as a cut-off for putative Hb enhancers (red line). Scores of the training enhancer sets are indicated in red. D) Distribution of scaled summary Hb scores for the genome-wide scans (gray) compared to the distribution of scaled summary Hb scores for the sequences which were classified as putative Hb enhancers by all classifiers (red).

Figure 3-5. Hindbrain enhancers can be accurately predicted from DNA sequence.



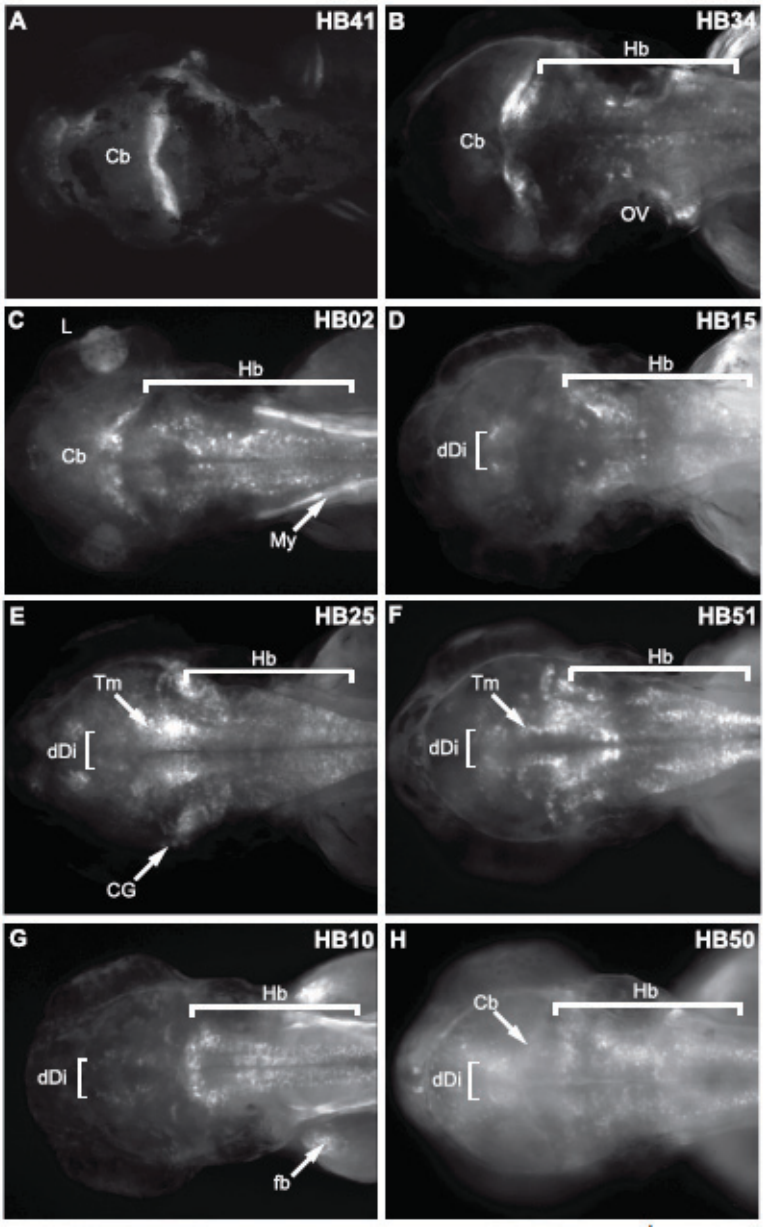
A) Area under the Receiver Operating Characteristic (ROC) curve (AUC) for three Hindbrain (Hb) enhancer classifiers trained on three highly overlapping datasets (enhancers with activity in the anterior Hb, posterior Hb, and whole Hb). AUC values range from 0.5 (random discrimination) to a theoretical maximum of 1. In a cross-validation setting our Hb classifiers obtained values of 0.89 (anterior Hb), 0.92 (posterior Hb), and 0.89 (combined Hb). B) Overlap among the top-scoring 5% Hb enhancer predictions produced by all three Hb classifiers. C) Fold-enrichment in 787 genes involved in Hb function in the neighborhood of positive predictions or putative Hb enhancers. Putative Hb enhancers were associated with the closest gene. *P*-values were computed using Fisher's exact test.

Figure 3-6. Experimental validation of tissue-specific enhancer candidates in transgenic assays in mice and zebrafish.



The *in vivo* validation rate of our computational Hb classifier, 5th column, trained on a small empirical dataset was comparable to those obtained with EP300 ChIP-Seq experiments in other brain tissues and limb (columns 3, 4 and 6) and exceeded that achieved in the heart by ChIP-seq methods or computational (columns 1 and 2).

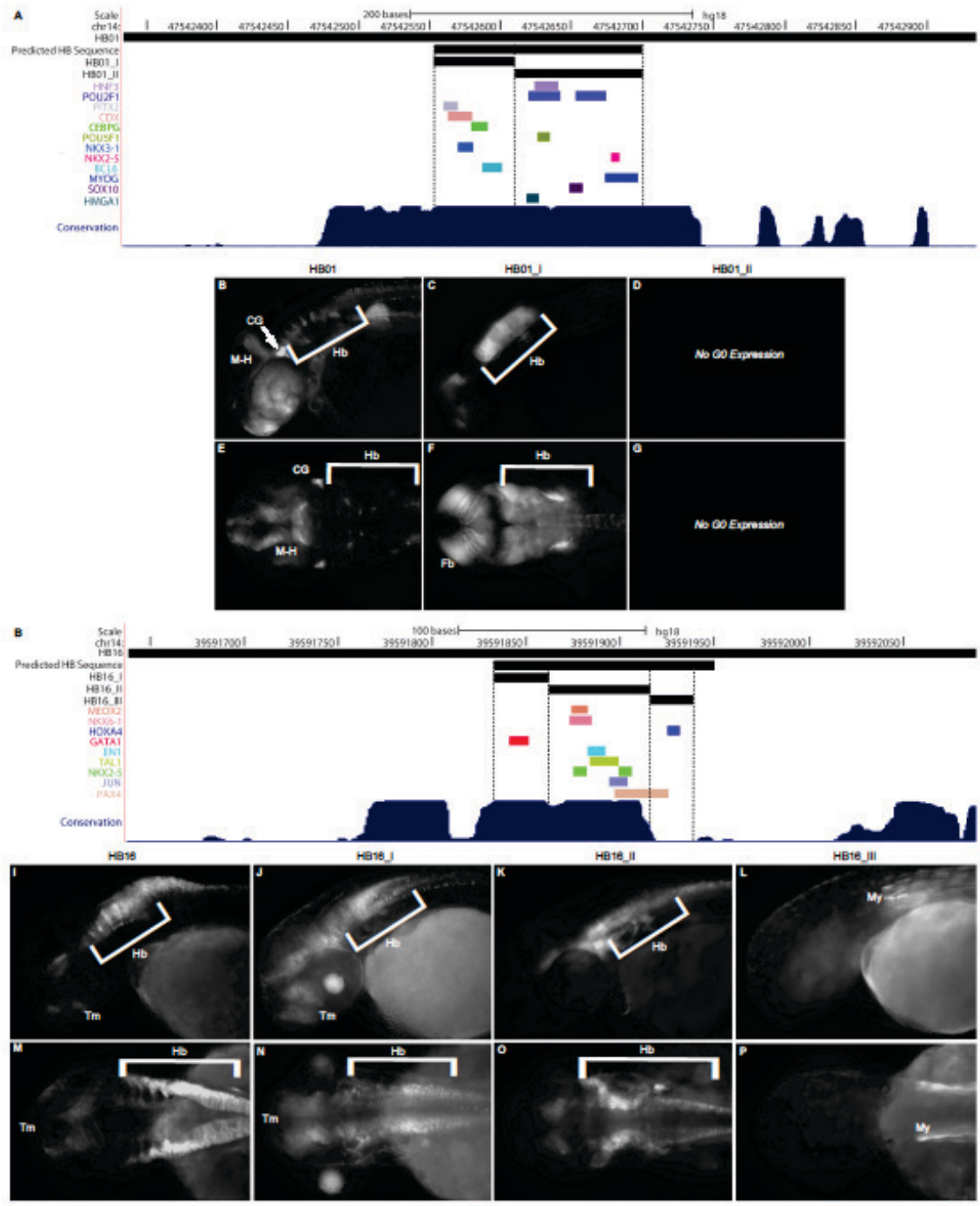
Figure 3-7. Predicted enhancers display pleiotropic expression patterns in the hindbrain.



A-H) GFP reporter expression from eight stable lines corresponding to Hb predictions showing expression across the Hb as well as in some non-Hb domains. Dorsal view images of EGFP reporter expression were taken at 3dpf.

Abbreviations: Cb – Cerebellum, CG – Cranial Ganglia, dDi – Dorsal Diencephalon, fb – Fin Bud, Hb – Hindbrain, L – Lens, M – Myotome, OV – Otic Vesicle, and Tm – Tegmentum.

Figure 3-8. Transcription factor motif clustering reveals functional sequence domains.



A and H) UCSC Genome Browser custom track showing injected construct, classifier predicted HB sequence, and fragments tested for Hb expression (black bars, top to bottom). Colored bars mark TFBS for various factors. B-G and I-P) EGFP reporter expression

observed with each sequence (lateral view, top; dorsal view, bottom). All images taken at 2dpf, anterior to the left. A) HB01 custom track with 2 subcloned fragments, B) Full length HB01, lateral view, C) HB01_I, lateral view, D) HB01_II, no G0 GFP reporter expression observed, E) Full length HB01, dorsal view, F) HB01_I, dorsal view, G) HB01_II, no G0 GFP reporter expression observed. H) HB16 custom track with 3 subcloned fragments, I) Full length HB16, lateral view, J) HB16_I, lateral view, K) HB16_II, lateral view, L) HB16_III, lateral view, M) Full length HB16, dorsal view, N) HB16_I, dorsal view, O) HB16_II, dorsal view, P) HB16_III, dorsal view. Abbreviations: CG – Cranial Ganglia, Fb – Forebrain, Hb – Hindbrain, L – Lens, M-H – Midbrain-Hindbrain Boundary, My – Myotome, and Tm – Tegmentum.

CHAPTER 4

APPLICATION OF CHIP-SEQ TO CREATE A CELL-TYPE-SPECIFIC REGULATORY VOCABULARY

4.1 Introduction

In our continuing effort to annotate the genome for transcriptional regulation and create the most accurate and predictive catalogs of cell-type restricted enhancers we set out to establish ChIP-seq as a viable technique to identify enhancers in our lab in a conservation independent, cell-type restricted, and genome-wide manner. For our early experiments we chose EP300, a transcriptional co-activator and histone acetyltransferase (Eckner et al., 1994; Vanden Berghe et al., 1999; Holmqvist and Mannervik, 2013), and the histone modification H3K4me1 as our proteins of interest. Both proteins are highly conserved, have been shown to be associated with enhancers in many cell types and tissues, and antibodies that have been used extensively for ChIP-seq are readily available (Visel et al., 2009, Blow et al., 2010).

We first focused on a mouse melanocyte cell line, melan-Ink4a-Arf, that was derived from epidermal melanocytes of an early post-natal *Ink4a/Arf* null pup on a C57BL/6J background (Bennett et al., 1987, Sviderskaya et al., 1997). Knock-out of *Ink4a/Arf* prevents senescence of the cells in culture, but has no other noticeable effects on phenotype. Melan-Ink4a-Arf cells are darkly pigmented, genomically stable, closely resemble primary melanocytes and are easily expandable to the numbers needed for ChIP-seq (100-200 million cells per replicate). As this was largely the thesis project of a prior student, I will only briefly touch on the data that led to my further work.

Using this method we identified a pattern of H3K4me1 peaks flanking EP300 peaks at known melanocyte enhancers. After applying this filter to the remaining data we

discovered 2,489 putative mouse melanocyte enhancers (Gorkin et al., 2012). This method achieved a high validation rate both *in vitro* (86%; 43/50) and *in vivo* (70%; 7/10). The *in vivo* validation rate was especially promising since reporter expression from the positively scored sequences was seen solely in melanocytes. Due to this high success rate I attempted to apply the same method to purified neuronal populations.

In this chapter I will describe my early attempts at applying this ChIP-seq-based enhancer catalog identification strategy to a significantly smaller population (~4 million) of primary rat cortical neurons. Based on my work we determined that EP300 ChIP-seq fails with low cell number so I next investigated how well H3K4me1 identifies putative enhancers in the absence of EP300. I found that H3K4me1 could be used to identify enhancers in smaller cell populations, albeit with lower validation rates, in the absence of EP300 ChIP-seq data. Furthermore, this rate is increased when H3K4me1 peaks are filtered for conservation.

4.2 Results

4.2.1 H3K4me1 ChIP-seq in reduced cell numbers identifies regions of putative functional significance across the genome

Establishing a large population of purified neuronal subtypes is quite difficult as neurons are not generally actively dividing and neuronal precursors that are dividing must be differentiated by the addition of chemicals such as retinoic acid (SH-SY5Y) or butyric acid (MN9D, Rick et al., 2006). More recently, embryonic stem cells or pluripotent stem cells have been differentiated by the addition of specific transgenes (Friling et al., 2009; Caiazzo et al., 2011; Addis et al., 2011) in an attempt to achieve a progressively more homogenous neuronal population. However, the resulting populations may have limited utility due to differences in physiological characteristics and expression profiles between *in vitro* derived

neurons and their *in vivo* developed counterparts and questions of lineage stability (Rick et al., 2006; Balasooriya and Wimalasena, 2007; Holmberg and Perlmann, 2012). Furthermore, the immortalized cell lines that are most easily grown and maintained in culture (SK-N-SH and SH-SY5Y) are derived from metastatic neuroblastomas and therefore are made up of a mixed population of cells that have limited utility for study of specific neuronal subtypes (Biedler et al., 1973; Ross et al., 1983). Unable to find an adequate transformed cell line, we sought to obtain the largest population of primary neurons possible to complete ChIP-seq for EP300 and H3K4me1.

We collected approximately 4 million embryonic derived rat cortical neurons using established protocols (Gary and Mattson, 2001; Gary et al., 2003; Gary et al., 2012) and completed ChIP-seq as done previously in the melan-Ink4a-Arf cell line (Gorkin et al., 2012) for EP300 and H3K4me1 in duplicate, collecting an input sample from each experiment. Unexpectedly, sequencing reads from EP300 mapped to the rat genome (build Rn4) very poorly (~3%; Table 4-1). H3K4me1 ChIP-seq reads mapped significantly better with an average of 32% of reads mapping between two replicates (Table 4-1). The percentage of mapped reads for input samples were also relatively low, with an average of about 62% of reads mapping (Table 4-1). However, this result is somewhat expected due to the poor annotation and high proportion of gaps in the Rn4 build of the reference rat genome (Gibbs et al., 2004). Unfortunately, this does not explain the significant difference between the percentage of EP300 and H3K4me1 mapped reads and further work was required to understand the discrepancy.

4.2.2 EP300 binding is dynamic and transiently found at putative enhancers

As previously described (chapters 1 and 2), sequence conservation can be used as an indication of function when looking at specific genomic loci, however it can also be extrapolated to the genomic context when examining ChIP-seq data. To better understand the sequencing results we investigated the overall sequence conservation of peaks that were called from our EP300 and H3K4me1 ChIP-seq (using MACS, Zhang et al., 2008 or CisGenome, Ji et al., 2008, respectively). We found that the H3K4me1 peaks and regions 100 to 1500 bp that were flanked by H3K4me1 peaks are enriched for conservation while EP300 peaks are not enriched for conservation in this dataset (Figure 4-1). This is a troubling result as we, and others (Gorkin et al., 2012, Rada-Iglesias et al., 2011, Visel et al., 2009, Blow et al., 2010), have previously shown in multiple tissue and cell types, including forebrain and midbrain, that EP300 peaks are associated with peaks in conservation.

Upon further examination we found that the poor mapping success and lack of conservation of EP300 reads was due to insufficient ChIP enrichment and thereby sequencing of erroneous reads derived from sequencing adapter dimers and trimers. These results are corroborated by our previous work in melanocytes that indicated that saturation of H3K4me1 sequencing is achieved by approximately 10 million reads, while saturation for EP300 ChIP-seq requires an excess of 30 million reads (Gorkin et al., 2012). We attempted to circumvent this problem in rat cortical neurons by increasing the number of reads generated for EP300 relative to H3K4me1, however as it appears to be a problem first with the amount of ChIP enrichment and then with the sequencing depth this increase in read number did not help achieve saturation. We determined from this data that with the present technology, it is not possible to complete EP300 ChIP-seq in decreased cell numbers. However, the intimate nature with which histones interacts with DNA (Sheffield and Furey, 2012, Tsompana and

Buck, 2014) make ChIP-seq for histone modifications in low cell numbers an attractive alternative.

4.2.3 H3K4me1 ChIP-seq can be used to identify putative enhancers in the absence of EP300 signal

Next, we examined the ability of H3K4me1 ChIP-seq to identify enhancers genome-wide in the absence of EP300 data. Going back to our H3K4me1 melanocyte dataset (Gorkin et al. 2012), we randomly selected 10 H3K4me1 flanked regions, regardless of EP300 status. We then selected another 10 regions that were filtered for conservation by overlapping with PhastCons elements (PhastCons30way; Siepel et al., 2005). These regions were cloned into a luciferase reporter vector for an *in vitro* study of functional activity in cultured melanocytes. Although only a small number of sequences were successfully cloned and tested (n=7 unfiltered, n=8 Phastcons filtered), we found that the fraction of elements showing a more than three fold change above the empty vector control was increased in the Phastcons filtered set. Only 57% of unfiltered H3K4me1 flanked regions (4/7) displayed luciferase activity above this threshold, while 75% of Phastcons filtered H3K4me1 flanked regions (6/8; Figure 4-2) drove threefold higher activity. Importantly, the *in vitro* validation rate of regions selected for H3K4me1 flanked regions that overlap with EP300 peaks was still quite a bit higher with 86% (43/50, Gorkin et al., 2012). However, these data show that while H3K4me1 signal alone is not as effective at identifying putative enhancer regions as EP300, when combined with additional data, such as conservation, it is possible to increase its predictive power and may be used to identify putative enhancers in reduced cell populations.

4.3 Conclusions

We completed ChIP-seq for the transcriptional co-activator and histone acetyltransferase EP300 and the histone modification H3K4me1 in a relatively small population (~4 million) of purified primary rat cortical neurons with the hope of combining these two dataset, as we had previously published in Gorkin et al., to identify putative regulatory elements genome-wide. However, we show here that EP300 ChIP-seq performs extremely poorly in low cell numbers, with a very low proportion of mapped reads and no enrichment for conservation (Table 4-1, Figures 4-1 and 4-2). As a member of the mediator complex, EP300 does not interact directly with DNA, and instead functions by acting through transcription factors and histones. We predict that this property makes its interaction with DNA less stable and highly transient. These characteristics could explain how using reduced cell numbers could reduce our ability to establish high ChIP enrichment at any one location resulting in very low amounts of DNA that becomes overwhelmed by sequencing adapter dimers and trimers during sequencing library preparation.

Conversely, H3K4me1 sequencing reads mapped about 10 times better than EP300 and peaks called from H3K4me1 reads showed a marked increase in conservation (Table 4-1, Figure 4-1). This unanticipated result required that we further analyze the ability of H3K4me1 ChIP-seq to identify enhancers in the absence EP300 signal. To this end we selected 10 random H3K4me1 flanked regions from our previous work in melanocytes and 10 sequences with a Phastcons score above 0.15 for analysis of luciferase activity *in vitro*. This was based on the expectation that by filtering the H3K4me1 reads for conservation we would enrich for functional sequences in cases where we cannot apply an EP300 filter.

By applying a filter for conservation we were able to increase the proportion of sequences driving luciferase activity at least threefold above the empty vector control by

approximately 25% (Figure 4-2). This is an encouraging result that implies that we may continue to use H3K4me1 ChIP-seq in reduced cell numbers for the genome-wide identification of putative enhancers even when EP300 data is not available. H3K4me1 gives a much broader signal and is found more extensively across the genome than EP300 (3,622 EP300 peaks; 21,189 H3K4me1 flanked regions; >42,378 H3K4me1 peaks; Gorkin et al. 2012) so it is expected that when used on its own it will include a higher rate of false positives.

Although there continues to be a significant challenge in filtering out false positives from H3K4me1 ChIP-seq data, we believe that H3K4me1 ChIP-seq is an adequate predictor of functional activity. In combination with conservation, the predictive power H3K4me1 ChIP-seq is magnified. We continue to investigate new methods for extracting H3K4me1 peaks that are important for enhancer activity from background signals, including applying secondary and tertiary filters such as Phastcons score or ChIP-seq for additional histone modifications or transcription factors. Additionally, we predict that improved peak calling algorithms can further aid in improving our ability to extract signal from noise in H3K4me1 ChIP-seq. Despite these issues, we believe this data has shown that H3K4me1 can effectively be used for ChIP-seq in reduced cell numbers and may be applied to homogenous populations of specialized neurons for cell-type specific enhancer discovery.

4.4 Methods

4.4.1 Culture of rat cortical neurons

Primary rat cortical neurons were gifted by Devin S. Gary (Kennedy Krieger Institute, Center for Spinal Cord Injury) and were established from embryonic day 18 Sprague-Dawley rat cortices as described previously (Gary and Mattson, 2001; Gary et al,

2003, Gary et al. 2011). Pregnant females were euthanized and embryos removed. Cortices were dissected from pups, dissociated by trypsin treatment and tritured 15-20 times in Neurobasal media with B-27 supplement (Life Technologies) to break down tissue. Cell clumps were removed by passage through a 40-μm filter, and cells were plated on poly-D-lysine coated plates (10μg/ml, coated overnight) at a density of 75,000 cells/cm^2. Cortical neurons were maintained in neurobasal media and B-27 supplement for 5 days prior to ChIP.

4.4.2 EP300 and H3K4me1 ChIP-seq

ChIP for EP300 and H3K4me1 was completed as established in Gorkin et al. 2012 with minor changes in volume to account for differences in cell number. Cells were fixed with 1.1% formaldehyde for 10 minutes, which was then quenched with 125mM glycine. Fixed cells were then lysed in lysis buffer 1 (5mM PIPES, 85mM KCl, 0.5% NP-40, and 1x Roche Complete, EDTA-free protease inhibitor), and lysis buffer 2 (50 mM Tris-HCl, 10 mM EDTA, 1% SDS, and 1x Roche Complete, EDTA-free protease inhibitor). Cells were resuspended in 16.7mM Tris-HCl, 1.2mM EDTA, 167mM NaCl, 0.01% SDS, 1.1% Triton X-100, and 1x Roche Complete, EDTA-free protease inhibitor for sonication. DNA was sheared using a Bioruptor (Diagenode) with the following settings: high output; 30-sec disruption; 30-sec cooling; total sonication time of 40 min with addition of fresh ice and cold water to the water bath every 10 min.

Sheared DNA was incubated overnight for chromatin immunoprecipitation with anti-EP300 (Santa Cruz Biotechnology; sc-585) or anti-H3K4me1 (Abcam; ab8895). Antibodies were bound and immunoprecipitated by protein-G Dynabeads (Life Technologies) and collected by magnet at 4°. IPs were washed twice with low-salt wash buffer (0.1% SDS, 1% Triton X-100, 2mM EDTA, 20mM Tris-HCl, 150mM NaCl), twice with high-salt wash

buffer (0.1% SDS, 1% Triton X-100, 2mM EDTA, 20mM Tris-HCl, 500mM NaCl), and twice with LiCl wash buffer (250mM LiCl, 1% IGEPAL CA630, 1% deoxycholic acid [sodium salt], 1mM EDTA, 10mM Tris-HCl) and rinsed once with PBS (pH 7.4).

Beads were resuspended in elution buffer (50mM Tris, 10mM EDTA, 1% SDS) and incubated overnight at 65^0C. 2 volumes of elution buffer was also added to the input fraction and it was incubated overnight at 65^0C. Samples were treated with RNase A (0.2ug/ul) for 1 hour at 37^0C, and then proteinase K (0.2ug/ul) for 2 hours at 55^0C. DNA was purified by standard phenol-chloroform extraction, and resuspended in 10mM Tris-HCl. Two biological replicates were performed for each antibody, with each replicate consisting of a ChIP sample and an input (pre-IP) sample. Libraries were prepared using a standard TruSeq DNA sample prep kit (Illumina) and sequenced on an Illumin GA2 obtaining 100 bp reads. All four input samples and both H3K4me1 samples were pooled in one lane, while both EP300 samples were pooled in a second lane to increase read numbers.

4.4.3 Mapping and peak calling

Reads were mapped using Bowtie2 default settings to rat genome build 4 (Gibbs et al. 2004). EP300 peaks were called using the Model-based Analysis for ChIPseq (MACS) algorithm (Zhang et al. 2008). H3K4me1 peaks were called using CisGenome (Ji et al. 2008) because it tends to call separate peaks corresponding to each apex of the bimodal distribution of H3K4me1 signal flanking enhancers, whereas MACS tends to call the entire bimodal distribution as a single peak. The Two Sample Peak Calling option in CisGenome was used to call peaks from both replicates concurrently and produce a single set of output files. Default settings were used for both peak callers, except that 'half window size W' was set to 4 for CisGenome.

4.4.4 PhastCons scores

Average PhastCons score plots (Figure 4-1) were generated with the Conservation Plot tool as part of the Cistrome Analysis Pipeline using interval files of H3K4me1 peaks, regions 100 to 1500 bp apart flanked by H3K4me1 peaks, and EP300 peaks (Liu et al. 2011).

4.4.5 Luciferase assays

H3K4me1 flanked regions were selected from our previously published ChIP-seq dataset in the melan-a cell line (Gorkin et al. 2012) available on the NCBI Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE38498. Regions flanked by H3K4me1 were further filtered by overlapping with PhastCons (phastCons30way) elements obtained using the UCSC Table Browser (Karolchik et al. 2004). Sequences were PCR amplified from mouse genomic DNA (Promega) and TA-cloned into pCR8/GW/TOPO (Life Technologies). Constructs were then Gateway/LR cloned (Life Technologies) into a luciferase reporter construct containing the firefly luciferase gene downstream from a minimal E1B promoter (Antonellis et al. 2006). Melan-a cells were plated in 24-well plates (40,000 cells/well) in biological triplicate and transfected next day with 400 ng of luciferase reporter and 8 ng of pCMV-RL Renilla expression vector (Promega) using 2 mL Lipofectamine 2000 per well (Life Technologies). Cells were lysed at 48 hours post-transfection and assayed with the Dual-Luciferase Reporter Assay System (Promega) using a Tecan GENiosPro Microplate Reader (Tecan Group).

4.5 Tables: Chapter 4

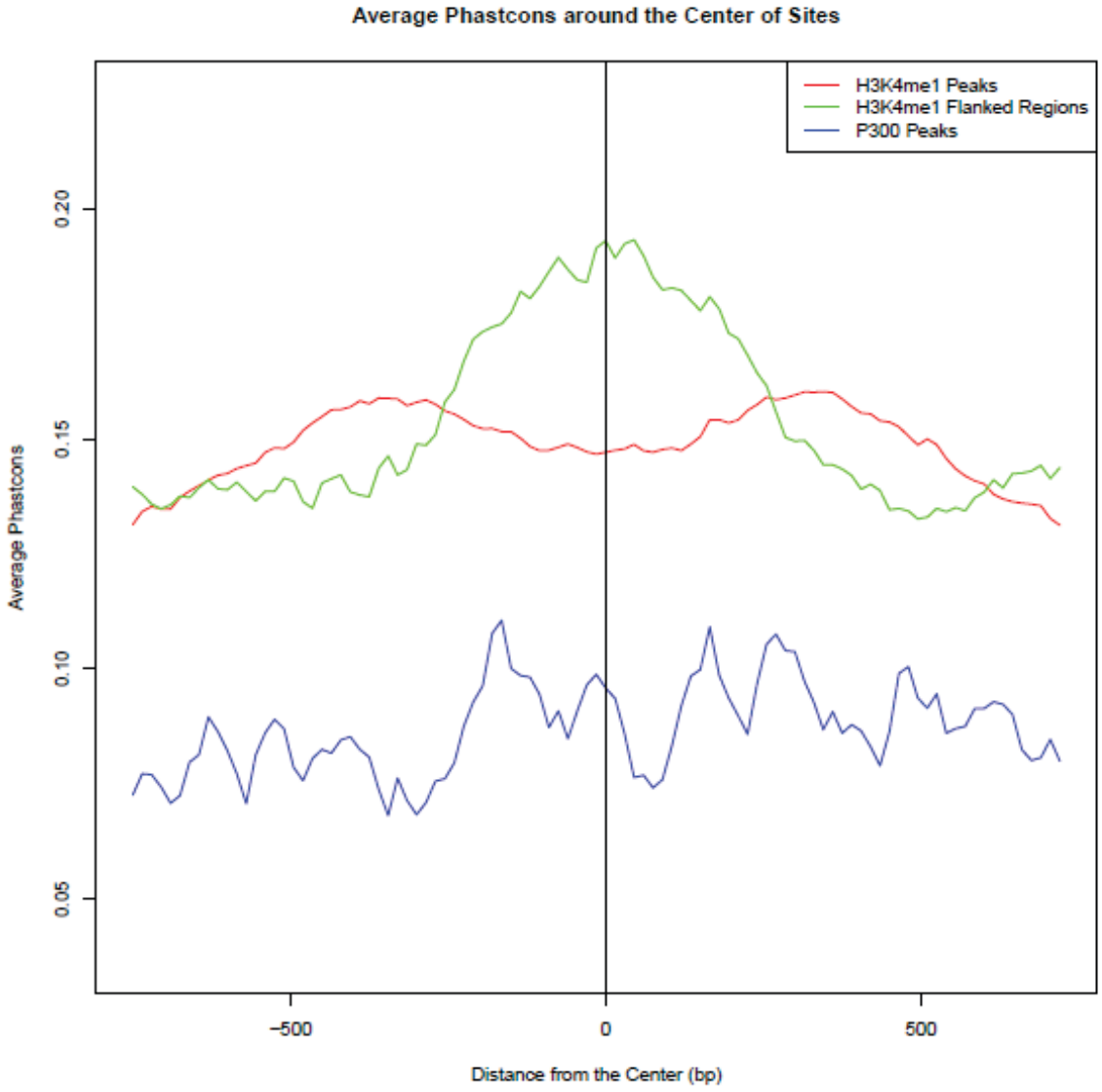
Table 4-1. Fraction of mapped reads from EP300 and H3K4me1 ChIP-seq replicates.

Sample	Total reads	Uniquely mapped reads	Percent uniquely mapped reads
Input Expt 1	18,445,970	10,969,042	59.47
H3K4me1 Expt 1	15,489,496	4,390,210	28.34
Input Expt 2	15,891,197	10,235,526	64.41
H3K4me1 Expt 2	18,552,435	6,777,966	36.53
Input Expt 3	16,337,212	10,008,333	61.62
EP300 Expt 3	71,975,408	2,444,601	3.40
Input Expt 4	24,077,331	14,951,967	62.10
EP300 Expt 4	43,377,370	1,532,905	3.53

Reads per experiment sequences on an Illumina HiSeq GA2. Sequences were mapped to the Rn4 buiied of the rat genome using Bowtie 2.

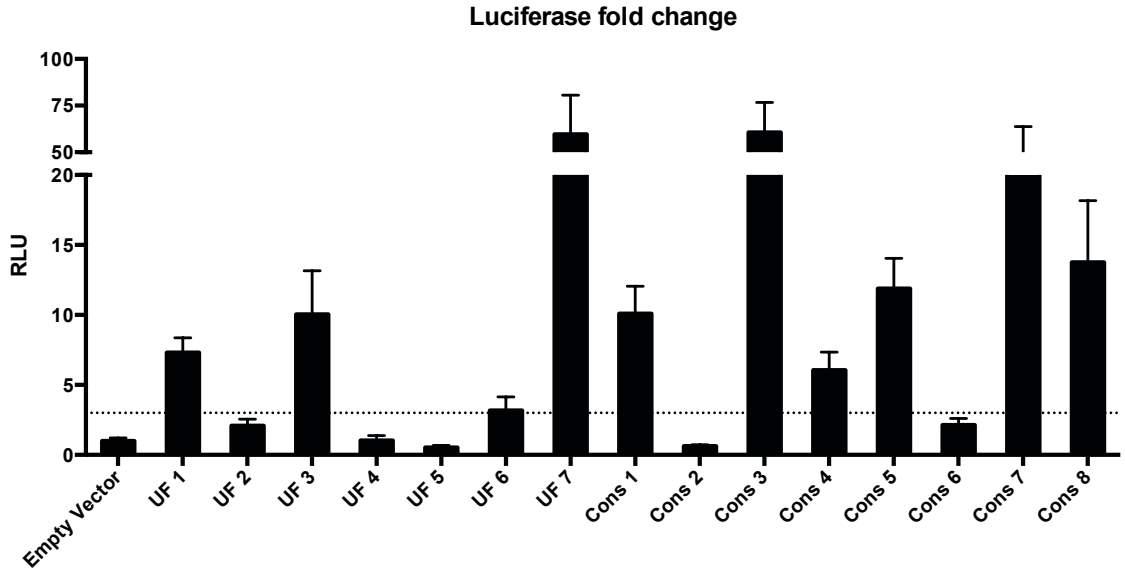
4.6 Figures: Chapter 4

Figure 4-1. H3K4me1 flanked regions from rat cortical neurons are enriched for conservation while EP300 peaks are not.



H3K4me1 peaks (red line) called by Cisgenome show slight peaks in conservation on either sie of the center, while regions that are flanked by H3K4me1 peaks (100-1500 bp apart; green line) show a spike in conservation at the center of sites implying functionality. However, EP300 peaks called by MACS (blue line) show no enrichment for conservation.

Figure 4-2. H3K4me1 flanked regions filtered for conservation show increased luciferase activity in vitro compared to regions not filtered for conservation.



57% of regions flanked by H3K4me1 peaks (100-1500 bp apart; UF 1-7) selected at random drive luciferase activity at least 3 fold higher (dotted line) than an empty vector control. Regions filtered for overlap with PhastCons elements (Cons 1-8) drive luciferase activity greater than 3 fold above empty vector control in 75% of cases.

CHAPTER 5

ESTABLISHMENT OF SMALL-SCALE CHIP-SEQ FOR THE ANALYSIS OF DOPAMINERGIC NEURONAL POPULATIONS *EX VIVO*

5.1 Introduction

The end goal of this thesis had always been to identify enhancers that are active in dopaminergic neurons and that may play a role in human variation and disease pathogenesis. Upon determining the value of H3K4me1 ChIP-seq in reduced cell populations we have shown in chapter 4 that this is a viable tool for enhancer discovery in neuronal subtypes. However, dopaminergic neurons are especially difficult to obtain in large numbers as they make up such a small proportion of the brain (Hitzemann et al., 1993).

While trying to identify the best cell substrate to complete H3K4me1 ChIP-seq we became aware of the Human Epigenome Atlas, a resource produced and maintained by the NIH Roadmap Epigenome Consortium (Bernstein et al., 2010). The Human Epigenome Atlas provides a valuable tool to researchers by making publicly available hundreds of ChIP-seq, RNA-seq, and DNaseI-seq datasets from human embryonic stem cells, as well as fetal and adult human tissues. We sought to investigate the adult brain tissues profiled by H3K4me1 from this expansive atlas. At least two replicates of H3K4me1 ChIP-seq, and an input sample, were available for port-mortem micro-dissected brain tissue from the anterior caudate, hippocampus, temporal lobe and, most significantly for my study, the *substantia nigra*. We investigated the ability of these highly biologically relevant datasets from human tissue to predict cell-type restricted enhancers genome-wide.

H3K4me1 ChIP-seq in these brain sub-regions was able to detect sequences that may function in gene regulation, successfully identifying regions that are enriched for

conservation and are often located near the top 1500 most highly expressed genes in the *substantia nigra*. However, using machine learning we discovered that although significant k-mers with predicted biological relevance were identified, the auROC of our datasets could not reach higher than 0.70 irrespective of the filters that were applied. We tested a small number (*substantia nigra* unique and conserved (SN – U, Cons), n=10; *substantia nigra* unique, conserved and contains motifs for FOXA2 and NR4A2 (SN – U, Cons, FN), n=10) of elements in zebrafish to examine reporter expression but saw very little expression in the CNS in either data set, only 3/10 for SN – U, Cons and 5/10 for SN – U, Cons, FN. We hypothesize that the auROCs and low validation rate is due to the heterogeneity of the tissues in which the ChIP was completed. Micro-dissected tissues are made up of a variety of cell types, which introduces noise to the ChIP resulting in sub-optimal enrichment. After observing such low enrichment and *in vivo* validation we determined that these were not ideal substrates and we must pursue a more homogenous population of dopaminergic neurons in order to discover an enhancer vocabulary for dopaminergic neurons.

To this end, we obtained a transgenic mouse line from the Gene Expression Nervous System Atlas (GENSAT; Gong et al., 2003; Heintz 2004) that expresses EGFP under the regulatory control of the human tyrosine hydroxylase locus (TH; Tg(Th-EGFP)DJ76Gsat/ Mmnc; from the Mutant Mouse Regional Resource Core at UNC; Khaliq and Bean, 2010; Ibáñez-Sandoval et al., 2010). TH catalyzes the rate-limiting step in catecholamine synthesis from tyrosine to dopa (Nagatsu et al., 1964) and is characteristically expressed in dopaminergic neurons. Mutations in the coding region of TH are linked to autosomal recessive forms of dopamine deficiency called Segawa syndrome (OMIM #605407; Ludecke et al. 1995; Brautigam et al. 1999). Mouse Th is expressed by embryonic day 11.5 (E11.5) as neuronal precursors are undergoing terminal differentiation to dopaminergic neurons

(Prakash and Wurst, 2006). Wanting to collect as many EGFP positive cells as possible, we dissociated the ventral midbrain at E15.5, under the assumption that the majority of precursors should be differentiated by then. However, especially in early experiments this only resulted in about 1-2% of cells sorted from the ventral midbrain being EGFP positive (~12,500 cells per embryo; 125,000 per litter). The low number of cells retrieved requires further optimization of available ChIP-seq protocols in order to obtain sufficient enrichment of H3K4me1 bound chromatin.

At the time of this study, a number of low cell ChIP-seq protocols had been published (Goren et al., 2010; Adli et al., 2010; Adli and Bernstein, 2011; Gilfillan et al., 2012) presenting evidence that to accurate profiling of histone marks could be done in as few as 10,000 cells per ChIP. We set out to optimize these protocols in our lab starting first with the fixed ChIP (X-ChIP) protocol from Adli and Bernstein, and later examined the utility of the native ChIP (N-ChIP) from Gilfillan et al. The Adli protocol was similar to the one we used previously for ChIP-seq in the melan-a cell line and rat cortical neurons, however due to the small cell numbers it employs a random amplification step after DNA isolation to increase the amount of library DNA. On the other hand, N-ChIP does not use formaldehyde to fix proteins to DNA instead employing micrococcal nuclease (MNase) to cut between histones and relying on the tight association of nucleosomes with the DNA to retain the proximity of histone modifications with functional sequences. Ultimately, the combination of a slightly altered, reduced-volume, standard ChIP protocol currently used by the Adli lab with an optimized sonication procedure and a ChIP specific library preparation kit proved to be necessary for examining H3K4me1 marks in purified dopaminergic neurons.

Upon final optimization of ventral midbrain dissection, dissociation, and FACS we were able to retrieve approximately 500,000 ventral midbrain EGFP positive cells per litter of

Tg(Th-EGFP)DJ76Gsat embryos at E15.5. We collected 20,000 EGFP cells per sort for RNA isolation and qRT-PCR, showing that these cells are indeed dopaminergic. The remaining EGFP positive cells were split in half to complete 2 different ChIPs, one with H3K4me1 and one with H3K27ac. ChIP-seq for H3K27ac, a histone modification associated with active developmental enhancers (Creyghton et al., 2010; Rada-Iglesias et al., 2011), was completed to add another filter to the H3K4me1 data and aid in identification of functional sequences.

Using these methods we identified a catalog of putative dopaminergic enhancers. These sequences display increased conservation at the center of sites and are enriched near genes associated with neuronal function and processes. Regions for *in vivo* validation were selected based on their proximity to genes involved in dopaminergic processes or association with Parkinson's disease. These studies are ongoing and will be completed by another competent and reliable graduate student.

5.2 Results

5.2.1 H3K4me1 ChIP-seq from human brain tissue identifies putative dopaminergic neuron enhancers

Making use of publically available datasets for dopaminergic neurons we downloaded H3K4me1 ChIP-seq replicates for anterior caudate, hippocampus, temporal lobe and *substantia nigra* from the Human Epigenome Atlas. We re-called peaks from the mapped sequencing reads of each dataset and identified H3K4me1 flanked regions as described previously (Chapter 4; Gorkin et al., 2012; Table 5-1) to locate putative enhancer sequences for each brain region. To discover regions that may be necessary for specific activity in the *substantia nigra,* all H3K4me1 flanked regions identified in the other brain sub-regions that overlapped with *substantia nigra* flanked regions were removed. The remaining *substantia*

nigra sequences identified by H3K4me1 ChIP-seq were then dubbed *substantia nigra* unique dataset (Table 5-1). As a result, all sequences that were identified in at least one other brain sub-region studied were excluded from further analysis.

To better understand the characteristics of our *substantia nigra* H3K4me1 marked sequences we first examined the average conservation level of regions using the Conservation plot tool (Siepel et al. 2005) on the Cistrome Analysis pipeline (Liu et al. 2011). The average PhastCons score was increased at the center of sites of the entire *substantia nigra* dataset, and conservation could be further enriched by overlapping *substantia nigra* unique H3K4me1 flanked regions with PhastCons elements (Figure 5-1). Next, *substantia nigra* U, Cons regions are enriched near genes that are highly expressed in human dopaminergic neurons (Marei et al. 2011; Figure 5-2; black bars), relative to those that are lowly expressed (Figure 5-2; light gray bars) or genes randomly selected from the expression range (Figure 5-2; dark gray bars). These results imply that our catalog may play a role in gene regulation in dopaminergic neurons.

5.2.2 Microdissected brain tissue does not provide a homogenous cellular substrate for enhancer prediction

Next, we investigated the predictive power of these datasets by using machine learning k-mer-based SVM (Lee et al., 2011; Gorkin et al., 2012) to identify overrepresented k-mers. The top positively weighted k-mers identified include motifs for FOXA2 and NR4A2, transcription factors with known roles in dopaminergic neuron development and maintenance (Table 5-2; Ferri et al., 2007; Lin et al., 2009; Lee et al., 2010; Stott et al., 2013; Saucedo-Cardenas et al., 1998; Smits et al., 2003). Other positively weighted k-mers contain motifs for SOX family TFs, including SOX2 which is established as an important factor in

the maintenance of neural stem cells (Graham et al., 2003; Ferri et al., 2004; Liu et al., 2014). Factors with high negative weights include RORα (Table 5-2), a factor required for the differentiation and survival of cerebellar Purkinje neurons (Doulazmi et al., 2001; Serra et al., 2006; Serinagaoglu et al., 2007), as well as members of the HOX and GATA families which are involved in hindbrain segmentation and development of numerous non-neuronal cell types (Tsang et al., 1997; Tennyson et al., 1998; Ohnemus et al., 2001; Huang et al., 2009; Miguez et al., 2012). Therefore, biologically relevant explanations for the top k-mers can be hypothesized, as the positively weighted k-mers have well described roles in dopaminergic neurons, while negatively weighted factors appear to function in the development and differentiation of other cell types.

Although the highly weighted k-mers appear biologically relevant, auROCs for each k-mer-SVM classifier from the *substantia nigra* datasets were surprisingly low. Unfiltered *substantia nigra* H3K4me1 flanked regions achieved an auROC of only 0.597 while filtering *substantia nigra* elements for unique identifcation and conservation (U, Cons) modestly improved the auROC to 0.650 (Figure 5-3). We sought to improve the auROC of the classifier by filtering the *substantia nigra* U, Cons set even further sequences that contained both FOXA2 and NR4A2 motifs.

Since k-mers for these factors were among the most highly weighted it seemed a reasonable expectation that additional selection based on these motifs would identify a more homogeneous catalog of putative dopaminergic enhancers, however the auROC was improved only slightly to 0.677. These auROCs are quite low relative to the auROC of 0.912 achieved for EP300 and H3K4me1 overlap in melanocytes (Gorkin et al., 2012). Even when excluding EP300 peaks, the auROC for H3K4me1 flanked regions in melanocytes exceeded 0.70 and H3K4me1 flanked regions filtered for conservation achieved an auROC of 0.776

(Figure 5-3). We expected that the auROC of a high quality classifier would reach at least 0.70 since 0.50 describes random discrimination (Lee et al., 2012).

Despite the low auROCs we believed there was enough evidence, increased conservation (Figure 5-1), enrichment near highly expressed genes (Figure 5-2), and motifs for FOXA2 and NR4A2 among the top ranked k-mers (Table 5-2), to merit *in vivo* analysis of H3K4me1 flanked regions. We chose 10 regions from the collection of *substantia nigra* U, Cons as well as 10 sequences from *substantia nigra* U, Cons, with motifs for FOXA2 and NR4A2. Sequences to be tested in transgenic zebrafish were selected based on their proximity to loci associated with Parkinson's disease (Table 1-1). Both sets of elements attained very low rates of *in vivo* validation (*substantia nigra* U, Cons = 0.35; *substantia nigra* U, Cons with motifs for FOXA2 and NR4A2 = 0.50). For most elements tested, although mosaic expression in the CNS was observed, multiple founders could not be identified in the F2 generation. One notable exception to this is seen for +10.54 ARHGEF2 (Figure 5-4), which drives reporter expression in discrete populations of the CNS. Some of these populations appear to overlap with VMAT2:EGFP (a marker of monoaminergic neurons; Wen et al., 2007) marked cells in the diencephalon, forebrain and hindbrain (Figure 5-4, G-I). Overall, we were unable to validate publicly available H3K4me1 datasets for the *substantia nigra* downloaded from the Human Epigenome Atlas as substrates for enhancer identification.

We hypothesize that the heterogeneity of microdissected tissue obtained from this study introduce too many confounding signals to establish a good training set for enhancer identification in dopaminergic neurons. This result is in direct opposition to how the heterogenous training set of hindbrain elements performed in predicting hindbrain enhancers with high sensitivity but low specificity. In this case we had very low sensitivity and

specificity. Despite this failure, we still believe that the training set and classifier for dopaminergic neurons can be improved by using a homogeneous population of primary cells.

5.2.3 Transgenic mice containing a BAC encompassing the tyrosine hydroxylase locus provides a homogenous population of dopaminergic neurons for ChIP-seq

Although the tissues collected by the Human Epigenome Atlas were extremely biologically relevant they was not made up of pure population of neuronal subtypes. Obtaining a pure population, as we previously reported with melanocytes (Gorkin et al., 2012), will likely be the best way to remove all confounding signals and create the homogenous training set needed for enhancer discovery in dopaminergic neurons and predictions in the human genome. Towards this end, I decided to isolate a population of predicted dopaminergic neurons from Tg(Th-EGFP)DJ76Gsat mice.

Tg(Th-EGFP)DJ76Gsat mice are incompletely characterized (GENSAT; Gong et al., 2003; Heintz, 2004) however they have been shown to express EGFP in the ventral midbrain (Khaliq and Bean, 2010; Ibáñez-Sandoval et al., 2010). We dissected brains from E15.5 embryos, noting strong EGFP expression in the ventral midbrain as well as fainter expression in the striatum and hypothalamus of the forebrain (Figure 5-5 A). The ventral midbrain was removed and dissociated to a single cell suspension in preparation for FAC sorting. Dead cells were excluded during the sort by the addition of propidium iodide (PI) to label cells with damaged membranes (red). Two populations of cells were collected PI-, EFGP+ and PI-, EGFP- for RNA expression analysis (Figure 5-5, B).

We completed qRT-PCR for a number of genes that are characteristically expressed in neurons, Th, Slc6a3 (dopamine transporter), Pitx3, Foxa2, and Nr4a2, as well as Gad2, a marker of glutamatergic neurons, and Actb as a housekeeping control (Figure 5-5, C). There

was a marked enrichment in the PI-, EGFP+ cells for markers of dopaminergic neurons relative to the PI-, EGFP- population, particularly for Th, Slc6a3 and PItx3 (each showing >50 fold higher levels of transcript). Significantly, the non-dopaminergic transcripts did not show the same enrichment. This analysis shows that EGFP+ cells isolated *ex vivo* from the ventral midbrain of Tg(Th-EGFP)DJ76Gsat mice are indeed a good substrate for ChIP in dopaminergic neurons.

5.2.5 H3K4me1 ChIP-seq in FAC sorted dopaminergic neurons identifies a catalog of putative enhancers

After extensive optimization of embryonic brain dissection and dissociation we completed more optimization of the standard ChIP-seq protocol. After overcoming various problems related to minimizing cell and DNA loss, shearing consistency, DNA isolation and sequencing library preparation we were finally ready to complete ChIP-seq for H3K4me1 and H3K27ac. Despite each ChIP sample only containing 250,000 cells and the low amount of DNA retrieved from each ChIP, the mappability of sequencing reads was vastly improved over what we had observed in H3K4me1 ChIP-seq from 4 million rat cortical neurons (Chapter 4; Table 5-3), with 65-75% of reads uniquely mapping for each sample. However, we also observed a high proportion of PCR duplicate reads (~40-60%; Table 5-3) that must be removed during peak calling to avoid false positives (Leleu et al., 2010). This result is expected given the low amount of input DNA and number of PCR cycles used for library prep, but reduced the number of usable reads significantly.

Peaks were called for each ChIP replicate using MACS (Zhang et al., 2008) to identify regions of enrichment because CisGenome did not perform well in this instance. We examined the homogeneity and predictive power of each set of peaks using the k-mer-SVM

classifier (Lee et al., 2011). Default MACS parameters resulted in a relatively poor classifier, with auROCs ranging from 0.54 to 0.71 (Table 5-4). To find the best dataset we changed multiple MACS parameters including P-value, m-fold, shift size, and lambda. Peak sizes ranged from an average of 171 to 1110 bp depending on the parameters used (Table 5-4). Despite the low auROCs, an increase in average PhastCons score was seen for all samples (Figure 5-6) implying functionality. H3K4me1 peaks consistently display higher conservation than H2K27ac peaks, and conservation of peaks from replicate one (Figure 5-6, A) was not as high as replicate two (Figure 5-6, B). This is likely due to minor differences in DNA input and read numbers between replicates.

H3K4me1 peaks were refined to exclude coding regions using the UCSC Table Browser (Karolichik et al., 2004) to download RefSeq exons and the subtract tool in Galaxy (Blankenberg et al., 2010; Goecks et al., 2010), resulting in a catalog of 6. Conservation peaks were reduced, but still enriched, upon the removal of regions overlapping coding sequences, implying that a large fraction of peaks overlap with exonic sequences (Figure 5-7, A). However, H3K4me1 peaks were distributed into two clusters relative to the transcription start site (TSS): one collection of sequences in regions close to the TSS (0-5 kb) and a second set between 50-500 kb away from the TSS (Figure 5-7, B). Using the Genomic Regions Enrichment of Annotations Tool (GREAT; McLean et al., 2010) to identify the gene closest to each putative enhancer we found a marked enrichment for H3K4me1 peaks near genes that are expressed in the midbrain and nervous system as a whole (Table 5-5). Additional evidence pointing to functionality in the CNS includes enrichment of H3K4me1 peaks near genes that involved in cell processes associated with brain development (Table 5-6). Although not significant in the genome-wide analysis by GREAT, many H3K4me1 peaks were found near genes that are associated with Parkinson's pathogenesis in humans.

5.2.6 *Selection of sequences for in vivo validation*

We selected 20 regions from our catalog of 7574 putative enhancers for analysis in zebrafish. Regions were selected based on their proximity to genes expressed in the *substantia nigra* or linked to the development of Parkinson's disease in humans (Table 5-7). Some genes important in the development and maintenance of dopaminergic neurons had multiple H3K4me1 peaks near their genic locus (<500kb from the TSS or located in the gene's intron), such as Th, Otx2, and Park2. In these cases regions closest to the TSS were selected for testing *in vivo*. Sequences will be cloned into pTEA-cfosGW, injected into Tg(VMAT2:EGFP) zebrafish embryos and analyzed at the mosaic and F1 stages for expression in dopaminergic neurons. These studies are ongoing and progress will be reported as results are obtained.

5.3 Conclusions

In an effort to obtain and validate a catalog of dopaminergic enhancers we have studied cells from human and mouse using H3K4me1 as a marker of putative enhancers. Although we observed a large peak in conservation at the center of H3K4me1 marked regions (Figure 5-1) as well as enrichment near genes that are highly expressed in the *substantia nigra* (Figure 5-2), the auROCs obtained by k-mer-SVM were not especially high, with most in the 0.60 – 0.70 range (Figure 5-3 and Table 5-4). In vivo validation for H3K4me1 regions derived from human brain tissue was very low, barely reaching 0.5 when test sequences were rigorously filtered. We hypothesize that this poor validation rate is due to the heterogeneity of cell types resulting from tissue microdissection and the low n (n=1) of brains tested.

In order to obtain a more homogenous cell population we went to great care to dissect, dissociate to single cells, FAC sort the ventral midbrains of Tg(Th-EGFP) mice (Figure 5-5 A and B). The resulting purified population was greatly enriched for markers of dopaminergic neurons (Figure 5-5 C). We completed ChIP-seq for H3K4me1 and H3K27ac in these cells and obtained a catalog of putative dopaminergic neuron enhancers. These catalogs were enriched for conservation around the center of sites (Figures 5-6 and 5-7). H3K4me1 peaks were enriched near genes that were expressed in the nervous system, particularly in the *subtantia nigra* and midbrain (Table 5-5) as well as near genes involved in processes important for accurate brain development (Table 5-6).

Surprisingly, the auROCs for this pure population were not dramatically increased over what we had observed in the human tissue datasets. Attempts to increase the auROCs involved modification of a number of parameters (Table 5-4) but the maximum auROC reached was 0.71. This unexpected result could be explained by low sequencing coverage and high PCR duplicates, additionally the amount of ChIP DNA was low and difficult to quantitate because of the large size range of library fragments (200 – 600 bp). As a result, read numbers were not consistent across pooled samples, cluster generation on the Hiseq was not optimized out and the number of usable reads for each ChIP was lower than ENCODE suggested guidelines of >20 million unique reads per sample (Table 5-3; Landt et al. 2012). These issues would likely not be as significant in a standard ChIP-seq experiment but when using small cell numbers they are difficult to overcome.

Although we could attempt further optimization of ChIP conditions or complete more *in silico* validation of our datasets the only thing that will tell us the true value of our catalog is *in vivo* validation. Speculation about sequencing results or various computational predictions has its place but there is a point at which we must carry on with the analysis of

the data that we have acquired. The initial set of sequences chosen for validation represent a group that we predict to give the highest chance of driving expression in dopaminergic neurons, due to their location at dopaminergic and Parkinson's loci. It is possible that we will not see expression in the zebrafish midbrain due to its differences in structure from mammals. In this case we may chose to do future reporter analyses in mice.

5.4 Methods

5.4.1 *Mice and midbrain dissociation*

Hemizygous male Tg(Th-EGFP)DJ76Gsat mice were obtained from the Mutant Mouse Regional Resource Core (MMRRC) at the University of North Carolina, Chapel Hill. Transgenic mice were crossed with wild-type female Swiss Webster mice (Charles River Labs). Pregnant dams were euthanized by isofluorane overdose followed by cervical dislocation. Pups were removed from the abdomen, placed in ice cold DMEM and euthanized by decapitation. Brains were dissected from embryos using forceps and scalpel under a dissecting light microscope (Zeiss) and placed in ice cold HBSS without Mg^{2+} and Ca^{2+} (Quantity Biological).

Fluorescent brains were identified using a fluorescent dissection microscope. The ventral midbrain was isolated by removal of the cerebellum and forebrain and tissue not showing EGFP expression was trimmed off. Remaining EGFP positive ventral midbrains were dissociated using the papain dissociation from Worthington Biochemical (Cat no. LK003150; Heuttner and Baughman, 1986). Briefly, tissue was minced with a scalpel, resuspended in papain (20 U/ml) and incubated at 37º for 30 minutes. Tissue was triturated with a glass pipette to break apart clumps every 10 minutes. After 30 minutes tissue was

triturated extensively then passed through a 40μm filter. Papain was neutralized, cells spun down, and then resuspended in HBSS with Mg^{2+} and Ca^{2+} for FACS.

5.4.2 *FACS*

Cells were stained with propidium iodide (PI) to label dead cells. The single cell suspension isolated from Tg(Th-EGFP)DJ76Gsat mice was FAC sorted on a Beckman Coulter MoFlo Cell Sorter with 3 lasers (UV 355 nm, Blue 488 nm, and Red 633 nm). Parameters were chosen to obtain the largest population of EGFP+, PI- cells possible, gating to obtain dim EGFP as well as bright in the same tube. Two populations were sorted for RNA analysis, 20,000 cells each for EGFP+, PI- and EGFP-, PI-. The remaining EGFP+, PI- cells were sorted from the suspension for ChIP.

5.4.3 *Chromatin immunoprecipitation and sequencing library preparation*

DNA bound to H3K4me1 and H3K27ac was immunoprecipitated using a protocol similar to that used in chapter 4 for rat cortical neurons and published in melanocytes (Gorkin et al., 2012) with some slight modifications. All ChIP steps were completed in 1.5mL microfuge tubes. Immediately after FACS completion cells were spun down in the HBSS buffer and resuspended in 1% formaldehyde (Sigma) and 10% FBS (Gibco) in PBS (Gibco). Cells were fixed for 10 minutes rocking at room temperature, then formaldehyde was quenched with the addition of glycine to a final concentration of 0.125M. Cells were spun down and washed twice with 10% FBS in PBS. Cells were lysed in SDS lysis buffer (50mM Tris-HCl (pH 8.1), 10mM EDTA, 1% w/v SDS + Roche EDTA free Complete Protease inhibitors).

Samples were diluted in ChIP Dilution Buffer (0.01% w/v SDS, 1.1% w/v Triton X-100, 1.2mM EDTA, 16.7mM Tris-HCl (pH 8.1), 167mM NaCl + Roche EDTA free Complete Protease inhibitors), then sonicated using a Diagenode Bioruptor. Settings were high, 30-seconds disruption, 30-seconds cooling in 5 minute cycles. Every 5 minutes samples were vortexed and spun down to remove any foam, and the Bioruptor reservoir was refilled with ice-cold water. Ice was used to cool the water down, but no solid ice was allowed to remain in the water bath during sonication.

After sonication like samples were pooled and 25µL was taken as input. The remainder was split in half for ChIP with H3K4me1 and H3K27ac. 2µg H3K4me1 (Abcam, ab8895) or 5µg H3K27ac (Abcam, ab4729) were added for each ChIP and samples were incubated rotating overnight at 4º. Protein A+G magnetic beads (Life Technologies) were incubated for 1 hour to bind antibodies. Complexes bound to the beads were washed twice with low salt wash buffer (0.1% w/v SDS, 1% w/v Triton X-100, 2mM EDTA, 20mM Tris-HCl (pH 8.1), 150mM NaCl), twice with LiCl wash buffer (0.25M LiCl, 1% w/v NP40, 1% w/v deoxycholate, 1mM EDTA, 10mM Tris-HCl (pH 8.1)) and twice with TE wash buffer (10mM Tris-HCl (pH 8.0), 1mM EDTA). Washed beads and input samples were incubated overnight at 65º in elution buffer (10mM Tris-Cl, pH 8.0, 5mM EDTA, 300mM NaCl, 0.1% SDS). Supernatants were removed and treated with 1mg/ml RNase A (Thermo Scientific) followed by 2mg/ml proteinase K (Thermo Scientific). DNA was isolated using the DNA Clean and Concentrator kit (Zymoresearch) as instructed. Library for sequencing was created using the Illumina Tru-seq ChIP Library prep kit (Illumina) as instructed. Samples were sequenced on an Illumina Hiseq 2500.

5.4.4 *RNA isolation and analysis*

RNA was isolated using an RNeasy kit (Qiagen) and cDNA was reverse transcribed using the Superscript III First Strand Synthesis kit (Life Technologies). Quantitative RT-PCR was completed using Power SYBR Green (Applied Biosystems) on a ViiA7 Real-Time PCR System (Applied Biosystems / Life Technologies).

5.4.5 *Mapping and sequencing analysis Peak Calling, k-mer-SVM, PhastCons, GREAT*

Reads were mapped to the mouse genome (UCSC mm9) using Bowtie2 (Langmead and Salzberg 2012) and filtered for a quality score of at least 20 to remove reads that map to more than one location using SAMtools (Li and Handsaker et al. 2009). Peaks were called with MACS (Zhang et al. 2008) testing variations in default parameters for model, shiftsize, lamda, and mfold. PhastCons plots for peaks were made using the Conservation Plot Tool on Cistrome (Siepel et al. 2005; Liu et al. 2011).

The kmer-SVM Galaxy server was used to create classifiers, determine auROCs, and identify highly weighted kmers for each dataset (Fletez-Brant and Lee et al. 2013). A five-fold greater size null set was used for each analysis. The GREAT server (McLean et al. 2011) was used to associate H3K4me1 peaks with the nearest gene and identify overrepresented characteristics of those genes. We employed the basal plus extension setting, with proximal elements being 5kb upstream or downstream of the TSS, and distal being up to 1Mb away from the TSS.

5.4.6 *In vivo validation*

Sequences for in vivo validation were selected based on their proximity to genes expressed in dopaminergic neurons or involved in the pathogenesis of Parkinson's disease

(Singleton et al., 2013; Lin and Farrer, 2014). H3K4me1 peaks overlapping conservation or ChIP-seq completed in other mouse brain tissues (Shen et al., 2012) were prioritized for analysis. Zebrafish were maintained as previously described (Kimmel et al., 1995; Westerfield, 2000). Sequences were cloned as described in chapter 3; predicted enhancers were amplified by PCR from human genomic DNA and cloned using Gateway Technology (Invitrogen). PCR fragments were TA-cloned into the pCR8/GW/TOPO vector (Invitrogen) then LR cloned into the pT2cfosGWvector, containing TdTomato instead of EGFP, for injection into Tg(VMAT2:EGFP) zebrafish embryos. Embryos were screening using a Zeiss Lumar V12 Stereo microscope with AxioVision software (version 4.5). Fish were raised to sexual maturity when >10% of screened embryos drove reporter expression in the CNS. Adults were crossed with Tg(VMAT2:EGFP) fish to identify stable lines.

5.5 Tables: Chapter 5

Table 5-1. H3K4me1 ChIP-seq identifies regions with predicted enhancer function in neuronal subregions.

Region	Peaks	Flanked Regions
Anterior Caudate	47,206	14,384
Hippocampus	58,204	19,146
Temporal Lobe	44,326	12,262
Substantia nigra	57,969	19,387
Substantia nigra – U, Cons	N/A	3596
Substantia nigra – U, Cons, with FOXA2 and NR4A2 motifs	N/A	653

Number of H3K4me1 peaks and flanked regions (100-1500 bp apart) indentified in each brain sub-region. Abbreviations: Cons – Conserved, U – Uniquely found in the *substantia nigra* dataset.

Table 5-2. Most significant k-mers in H3K4me1 flanked regions from substantia nigra unique, conserved dataset.

K-mer	Weight	Predicted TF	P-value
ACAAAG	2.39	Sox10	6.22E-05
		Sox12	1.12E-04
		Sox4	1.46E-04
		Sox11	1.91E-04
		Sox2	1.19E-03
		Pou5F1	2.39E-03
		Foxa2	5.06E-03
AAGGCC	2.22	HNF4A	2.24E-03
		NR4A2	2.28E-03
		Esrrb	7.19E-03
		PPARG::RXRA	1.11E-02
AATAAC	2.10	Foxa2	7.28E-04
		Tlx2	4.02E-03
ACCTGG	-2.91	ESR1	2.30E-04
		sna	4.74E-04
		Tcf3	1.31E-03
		ZEB1	2.21E-03
		RORα	5.85E-03
AGGTAA	-2.08	Hoxa6	9.88E-04
		Hoxa2	2.97E-03
		FOXF2	4.30E-03
GCATCC	-2.03	Spdef	6.75E-04
		Gata1	1.46E-03
		Ddit3::Cebpa	6.46E-03
		GATA2	6.69E-03
		ETS1	1.06E-02

Kmers identified from the substantia nigra – U, Cons classifier with weights greater than |2| and their predicted TFBSs.

Table 5-3. Fraction of H3K4me1 and H3K27ac reads from ChIP-seq in ex vivo isolated DA neurons mapping to mouse genome.

Sample	Total Reads	Reads mapped (mm9)	Uniquely mapped reads	Percent PCR duplicates
Input, R1	35,329,186	34,131,183	23,661,771	43
H3K4me1, R1	33,129,925	32,086,241	21,902,851	44
H3K27ac, R1	13,911,455	13,499,147	9,179,551	44
Input, R2	84,234,004	81,144,208	56,016,424	51
H3K4me1, R2	20,380,703	19,602,049	14,937,026	62
H3K27ac, R2	10,838,276	8,404,084	7,650,174	44

Sequencing completed on Illumina HiSeq GA2. Reads were mapped to the mm9 build of the mouse genome using Bowtie2. Percent PCR duplicates determined by MACS.

R1 – Replicate 1, R2 – replicate 2.

Table 5-4. Modification of peak calling parameters changes number of peaks and kmer-SVM auROC.

Sample	Rep	MACS options	Build Model	Dynamic Lambda	Peaks	auROC	average peak size
H3K4me1	1	nomodel; nolambda	n	n	752	0.594	1110.1
H3K27ac	1	nomodel; nolambda	n	n	839	0.543	853
H3K4me1	1	nomodel	n	y; ss=100	268	0.611	683.1
H3K27ac	1	nomodel	n	y; ss=100	455	0.540	573.4
H3K4me1	1	nomodel, ss=50	n	y; ss=50	690	0.541	310
H3K27ac	1	nomodel, ss=50	n	y; ss=50	1175	0.560	282.9
H3K4me1 no Input	1	nomodel; nolambda	n	n	752	0.594	1110.1
H3K27ac no Input	1	nomodel; nolambda	n	n	839	0.543	853
H3K4me1	2	nomodel; nolambda	n	n	6953	0.653	664.5
H3K27ac	2	nomodel; nolambda	n	n	4521	0.711	679.3
H3K4me1	2	nomodel	n	y; ss=100	5961	0.650	646.6
H3K27ac	2	nomodel	n	y; ss=100	4014	0.711	648.7
H3K4me1	2	nomodel, ss=50	n	y; ss=50	7970	0.619	310.3
H3K27ac	2	nomodel, ss=50	n	y; ss=50	3101	0.672	318.3
H3K4me1 no Input	2	nomodel; nolambda	n	n	6953	0.651	664.5
H3K4me1	2	mfold=10,30; llocal=20000	y	y; ss=100	3878	0.582	171.3
H3K4me1	2	mfold=10,30; llocal=10000	y	y; ss=100	3878	0.590	171.3
H3K4me1	2	mfold=5,30; llocal=10000	y	y; ss=100	6965	0.639	520
H3K4me1	2	mfold=5,30; llocal=20000	y	y; ss=100	6971	0.672	520
H3K4me1 Minus coding	2	mfold=5,30; llocal=20000	y	y; ss=100	7340	0.683	376.2

MACS options tested to obtain the best classifier from the mouse midbrain datasets. Further analysis and regions to be tested were completed on the last row.

Table 5-5. H3K4me1 peaks are enriched near genes expressed in the midbrain and central nervous system.

Term Name	Binomial Raw P-Value
TS28_substantia nigra	4.546E-10
TS28_dorsal raphe nucleus	1.319E-09
TS28_ganglion	2.008E-08
TS28_sympathetic ganglion	7.387E-08
TS28_hippocampal formation	4.913E-07
TS20_midbrain; lateral wall	6.038E-07
TS17_peripheral nervous system; autonomic; sympathetic	6.236E-06
TS26_brainstem	7.750E-06
TS21_telencephalon; mantle layer	9.484E-06
TS28_zona incerta	2.640E-05
TS13_midbrain-hindbrain junction	4.537E-05
TS17_rhombomere	6.550E-05
TS28_brainstem reticular formation	1.921E-04
TS15_future midbrain; roof plate	6.048E-04
TS28_ventral tegmental area	1.125E-03
TS28_pontine reticular formation	1.155E-03
TS21_optic nerve	1.900E-03
TS28_medial raphe nucleus	2.762E-03
TS28_substantia nigra pars compacta	3.039E-03

GREAT analysis of region-gene associations and expression locations from the MGI Expression database. P-Value is the Binomial Raw P-value.

Abbreviations: TS13 – 8.0-9.5 days post coitus (dpc); TS15 – 9.0-10.25 dpc; TS17 – 10.0-11.25 dpc; TS20 – 11.5-13.0 dpc; TS21 – 12.5-14.0 dpc; TS26 – 18 dpc; TS28 – postnatal.

Table 5-6. GREAT region-gene associations for H3K4me1 peaks are enriched for pathways and processes related to the nervous system.

Term Name	P-Value
Dendritic spine morphogenesis	1.019E-11
Midbrain-hindbrain boundary development	4.089E-11
Negative regulation of transforming growth factor beta receptor signaling pathway	4.933E-11
Dendritic spine development	4.990E-10
Negative regulation of epidermal growth factor receptor signaling pathway	9.626E-10
Rostrocaudal neural tube patterning	2.625E-09
Regulation of ARF protein signal transduction	2.847E-09
Negative regulation of epidermal growth factor-activated receptor activity	1.911E-08
Positive regulation of histone acetylation	4.620E-07
Substrate adhesion-dependent cell spreading	7.068E-06
Response to water	8.063E-06
Negative regulation of protein tyrosine kinase activity	1.374E-05
Response to food	2.742E-05
Parasympathetic nervous system development	2.797E-05
Preganglionic parasympathetic nervous system development	3.267E-05
Regulation of histone acetylation	3.275E-05
Stem cell division	1.223E-04
Enteric nervous system development	1.923E-04
Somatic stem cell division	2.894E-04
Cerebellar Purkinje cell differentiation	1.353E-03

Region-gene pathways and processes enriched in peaks from H3K4me1 ChIP-seq in *ex vivo* DA neurons. P-Value is the Binomial Raw P-value.

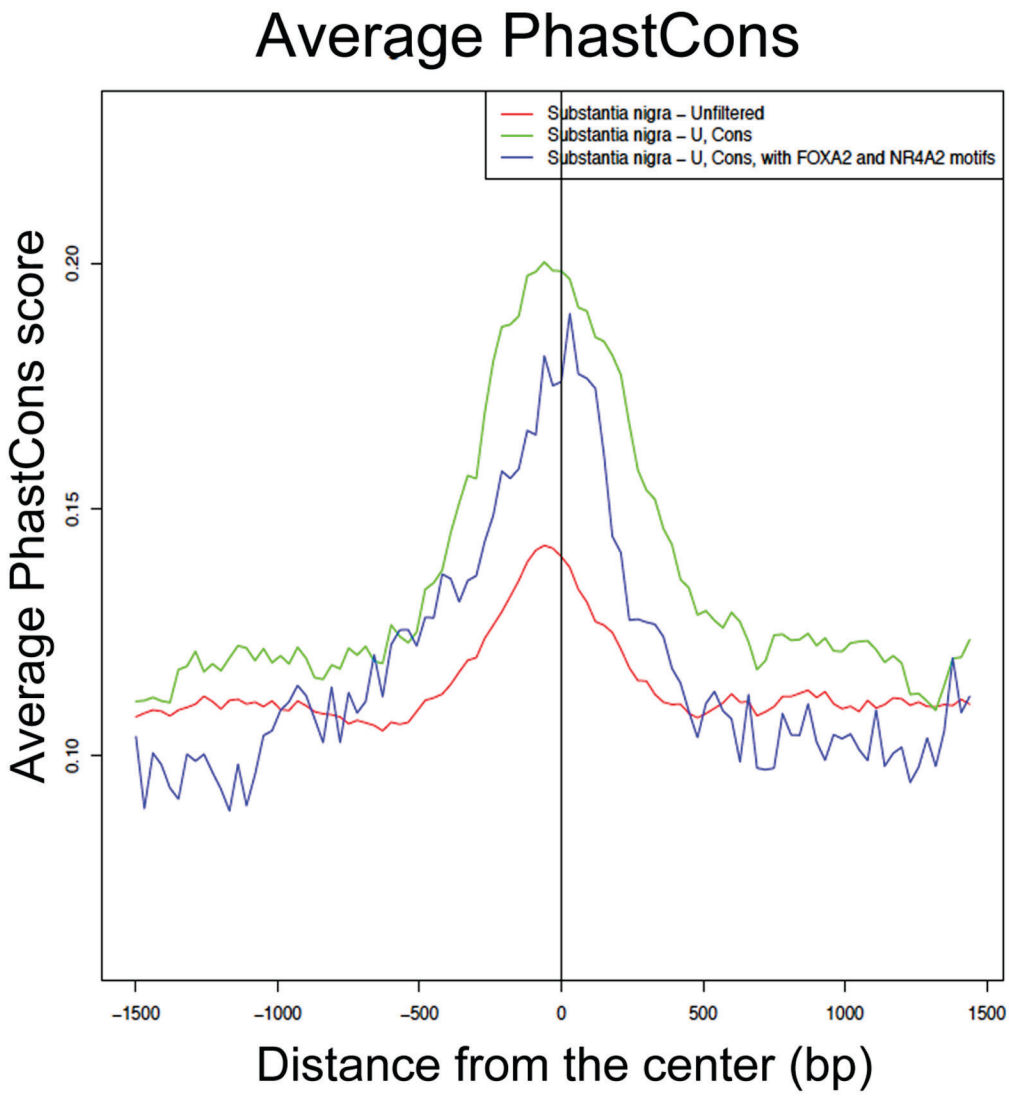
Table 5-7. Sequences selected from mouse VM ChIP-seq for in vivo analysis.

Construct	Coordinates of peak (mm9)	Genomic features
En1+213.76	chr1:122712847-122713355	Intergenic, En1-Celrr
Rxrg -99.82	chr1:169430008-169430440	Intergenic, Lrrc52-Rxrg, -190kb from Lmx1a
Nr4a2 prom	chr2:56967450-56967659	Proximal to Nr4a2 TSS
Sox2 -7.48	chr3:34541335-34541970	Sox-2OT, 3' of Sox2
Dnajc6 +0.631	chr4:101169901-101170380	Dnajc6 intron 1
Pink1 +0.783	chr4:137881453-137881747	Pink1 intron 1
Atp13a2 +25.96	chr4:140567690-140568369	Mfap2 intron, 25.96 to Atp13a2
Shh +2.77	chr5:28790860-28791435	Shh intron 1
Th- 7.70	chr7:150093577-150094369	Intergenic, Th-Ascl2
Th -18.01	chr7:150103805-150104945	Intergenic, Th-Ascl2
Th -18.55	chr7:150103805-150104945	Intergenic, Th-Ascl2
Slc6a3 -0.732	chr13:73673400-73673911	Intergenic, Lpcat1-Slc6a3
Otx2 -15.02	chr14:49293203-49293584	Otx2os1 intron
Wnt10b -2.31	chr15:98610890-98611378	Intergenic, Wnt10b-Wnt1, 8.9kb to Wnt1 TSS
Park2 +226.94	chr17:11260159-11260364	Park2 intron 2
Park2 +514.83	chr17:11548068-11548538	Park2 intron 4
Pax2 +51.28	chr19:44883125-44883519	Pax2 intron 6
Gbf1 +25.07	chr19:46252062-46252521	Gbf1 intron 3, 29 kb to Pitx3 TSS
Vax1 +40.40	chr19:59204088-59204568	Intergenic, Vax1-Slc18a2, 132kb to Slc18a2 TSS

Name of constructs to be tested in reporter assays *in vivo*. Genomic location of each H3K4me1 peaks, and relevant proximal features. Sequences were selected based on their proximity to genes linked or associated with parkinsonian phenotypes.

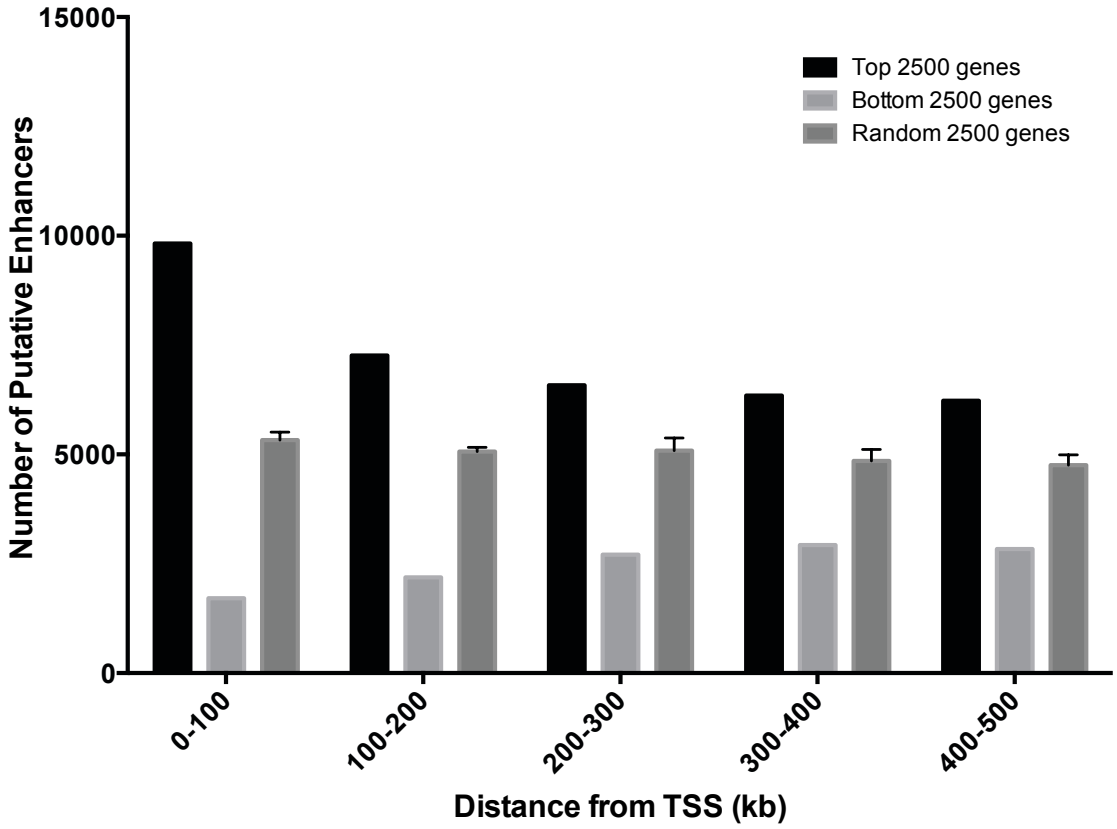
5.6 Figures: Chapter 5

Figure 5-1. Average PhastCons score is increased at the center of H3K4me1 flanked regions identified from human substantia nigra



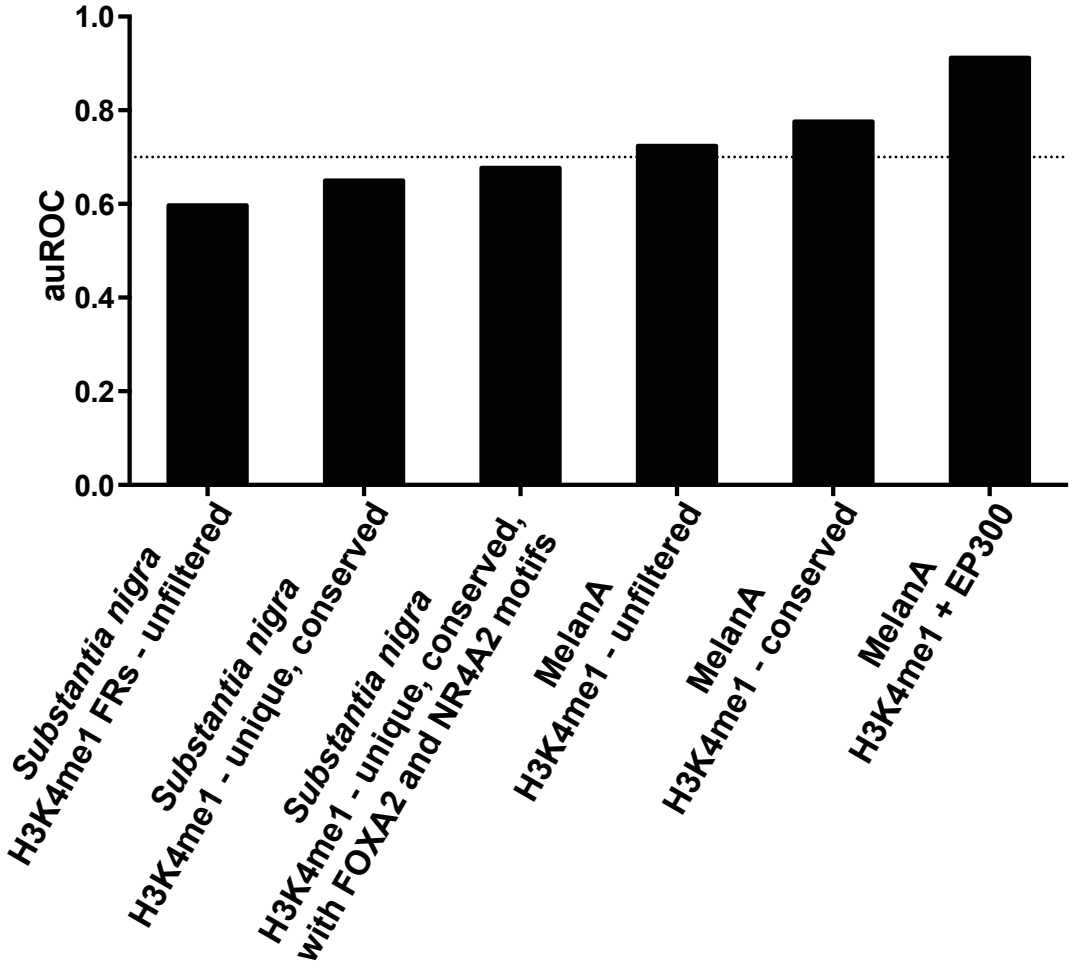
Average PhastCons score around the center of H3K4me1 flanked regions from *substantia nigra* datasets; all sequences identified by H3K4me1 ChIP-seq in the *substantia nigra* (Unfiltered; red); sequences overlapping conserved regions and not identified in any of the other brain sub-region datasets (U, Cons; green); Unique and conserved sequences that also have FOXA2 and NR4A2 motifs in them (U, Cons, with FOXA2 and NR4A2 motifs; blue).

Figure 5-2. H3K4me1 flanked regions from substantia nigra are enriched near the top 2500 expressed genes in dopaminergic neurons.



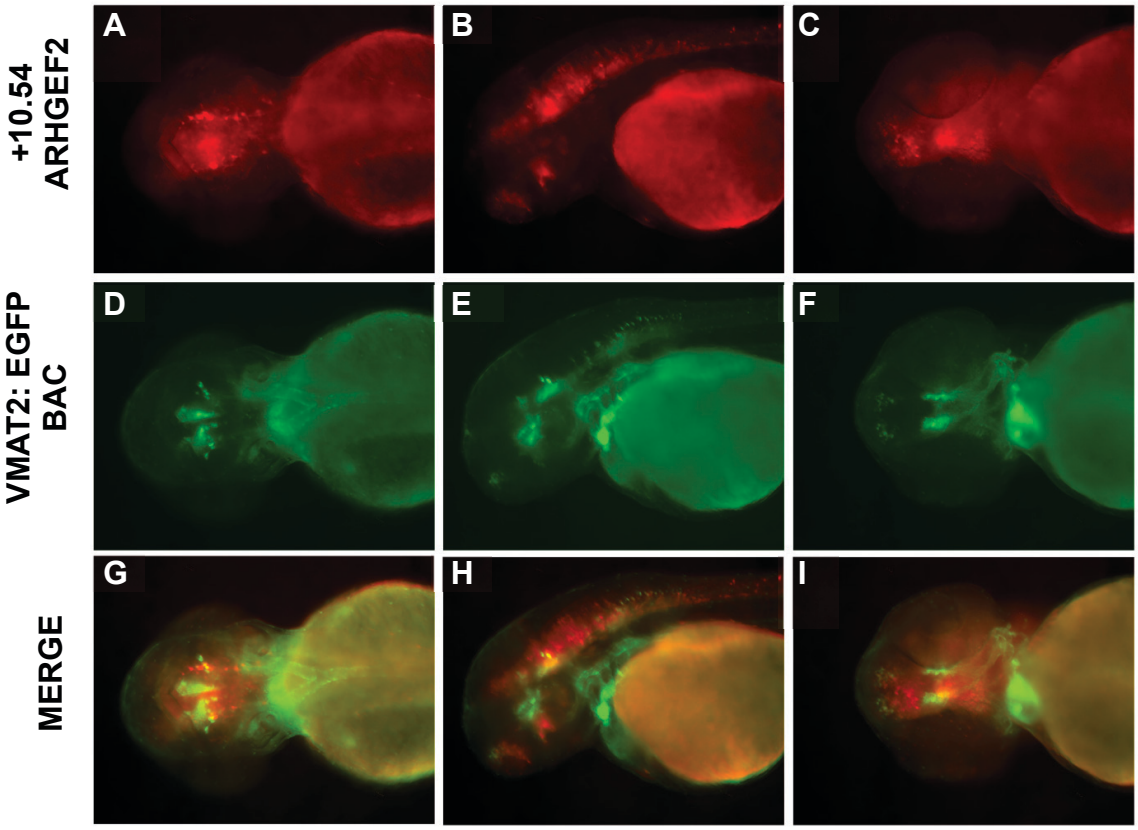
Putative enhancers identified by H3K4me1 ChIP-seq are enriched near the 2500 most highly expressed genes (black bar) in the substantia nigra. Regions are depleted near the 2500 lowest expressed genes (light gray), and equally distributed with regard to 2500 randomly selected genes (dark gray).

Figure 5-3. Classifiers for substantia nigra datasets do not perform as well as Melan-a ChIP-seq sets.



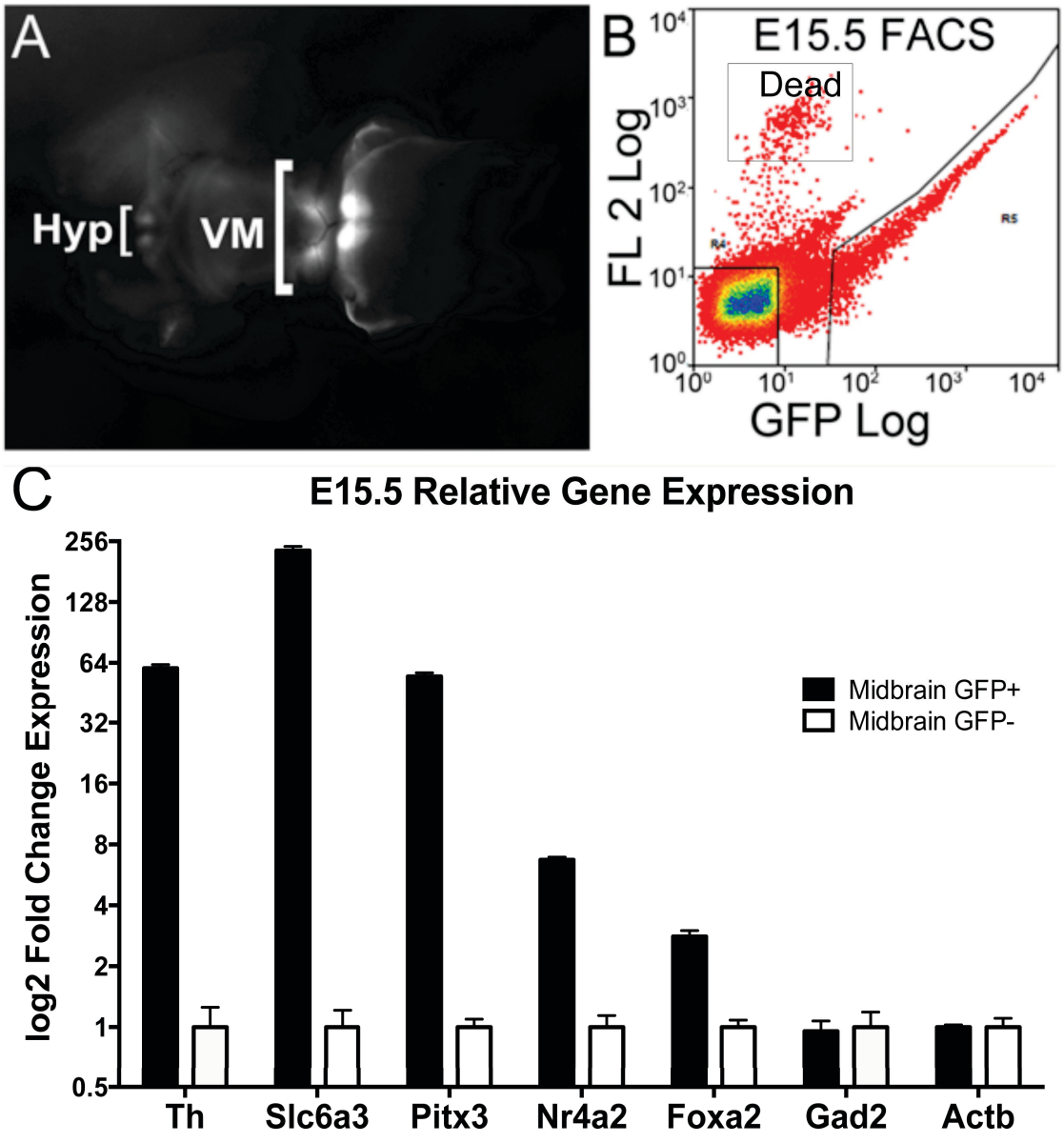
auROCs for classifiers from *substantia nigra* datasets increased with catalog refinement but still are significantly less than equivalent datasets from Melan-A cells. None of the *substantia nigra* catalogs reach 0.70 (dotted line).

Figure 5-4. +10.54 ARHGEF2 drives expression in discrete neuronal populations overlapping with dopaminergic neurons.



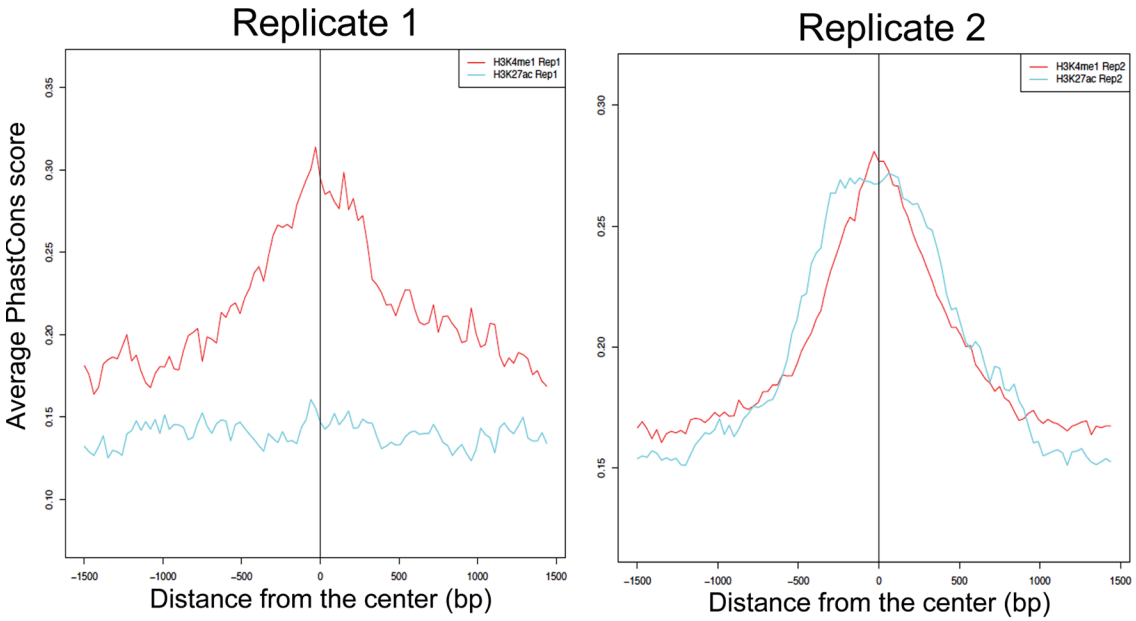
A representative stable line for +10.54 ARHGEF2, an element chosen based on its proximity (<100kb) to a SNP associated with Parkinson's disease and overlap with relevant TFBSs. A-C) show reporter expression of the +10.54 ARHGEF2 construct in CNS neurons. D-F) expression of the VMAT2:EGFP BAC marking monoaminergic neurons in the same embryo. G-I) show incomplete overlap of the putative enhancer with monoaminergic neurons.

Figure 5-5. Ventral midbrain EGFP+ cells are highly enriched for expression of dopaminergic neuron genes.



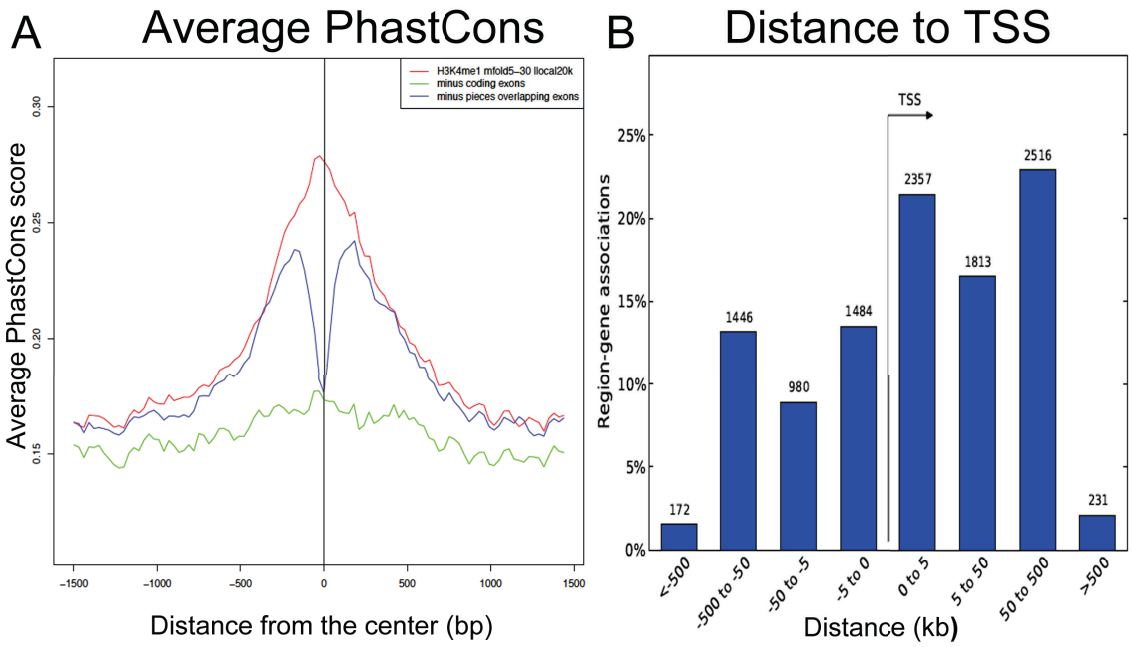
A) E16.5 Tg(Th-EGFP)DJ76Gsat brain, imaged ventrally with anterior to the left, posterior to the right. B) FACS plot of sorted neurons showing live, EGFP+ cells in R5, and EGFP- cells in R4. C) qRT-PCR shows that EGFP+ cells collected from the VM show high expression of DA neuron characteristic genes relative to EGFP- cells (normalized to Actb. Abbreviations: Hyp – hypothalamus, VM – Ventral Midbrain.

Figure 5-6. Average PhastCons score is increased at the center of H3K4me1 and H3K27ac peaks from ex vivo dopaminergic neurons.



The average PhastCons score is increased at the center of peaks from ChIP-seq in sorted mouse ventral midbrain neurons with H3K4me1 (red lines) in both replicates, and for H3K27ac (blue lines) in replicate 2.

Figure 5-7. Enrichment in conservation at H3K4me1 peaks is partially due to overlap with coding regions.



A) Unfiltered H3K4me1 peaks (red line) show a large peak in conservation at the center, but removal of all peaks overlapping with coding regions drastically reduces the conservation level. Sequences that only have the pieces overlapping exons (blue line) removed show a bimodal peak around the center. B) H3K4me1 regions show two peaks in region gene associations relative to the TSS with a pair of peaks between 0 and 5 kb away, as well as a second set 50 to 500 kb away.

CHAPTER 6

CONCLUDING REMARKS

6.1 Introduction

The preponderance of GWAS signals in noncoding sequences (Maurano et al., 2012) has implicated common variation in enhancers as a critical determinant of human phenotypes and disease. In this thesis I have used a variety of methods to uncover thousands of sequences predicted to regulate transcription in one more neuronal subtypes. The discovery of these enhancer sequences will be integral to furthering our understanding of brain development and disease. I began by using sequence conservation across species to identify non-coding regulatory sequences in a locus specific manner. This approach led to the identification of 22 new enhancers in the sequence flanking LMX1A (17) and LMX1B (5). Next, we sought to broaden our search space to all conserved non-coding regions throughout the genome. Using a machine learning approach based on a training set of 211 experimentally proven hindbrain enhancers we developed a primitive classifier that predicted over 40,000 sequences that may be active in the hindbrain. *In vivo* validation of this method revealed that 51/55 (93%) of tested elements directed CNS expression in mosaic zebrafish embryos, and 30/34 (88%) of these elements drove expression in the hindbrain of stable lines. The tested sequences gave highly pleitropic expression patterns and specificity to the hindbrain was low due to the heterogeneity of the training set. Despite this lack of specificity the sensitivity to detect sequences that will drive expression is still high, and this work has been critical in working towards the complete understanding of regulatory control in specific neuronal subtypes.

We next sought to increase the homogeneity of our training set by using a pure cell type rather than a tissue. This also involved the implementation of ChIP-seq to identify

putative enhancers. Based on previous work, we knew that ChIP-seq for EP300 and H3K4me1 could work well to identify enhancers genome-wide (Gorkin et al., 2012). After applying this same strategy to a much smaller population (4E6 vs 150E6) primary rat cortical neurons I found that H3K4me1 ChIP-seq still performed well, however EP300 did not. Upon further examination I determined that H3K4me1 ChIP-seq is an effective identifier of putative enhancers in the absence of EP300 data when peaks are also filtered for high conservation. Using luciferease assays we found that unfiltered H3K4me1 peaks drove luciferase activity 3 times greater than the empty vector control in 57% of constructs, while 75% of H3K4me1 elements that were filtered for conservation level showed at least 3 times greater activity. This result encouraged us to continue using H3K4me1 ChIP-seq in smaller cell numbers where EP300 would not work for the identification of putative enhancers.

While working to establish a pure population of DA neurons we set out to analyze publically available H3K4me1 ChIP-seq data from the Human Epigenome Atlas (Bernstein et al., 2010). I identified H3K4me1 flanked sequences in four human brain sub-regions: anterior caudate, hippocampus, temporal lobe, and the *substantia nigra*. With a goal of identifying putative enhancers specific to the *substantia nigra* I filtered this catalog to remove all sequences that were also found in any of the other brain sub-regions. This resulted in a catalog of 3,596 putative *substantia nigra* unique enhancers when also filtered for conservation status. I observed that these putative enhancers were more likely to be located near genes with high expression in the *substantia nigra* and they were enriched for TFBSs for FOXA2 and NR4A2, TFs important in DA neuron development. However, the catalog resulted in an unexpectedly low kmer-SVM auROC, unable to reach 0.70. We hypothesize that the microdissected tissue substrate used in the ChIP-seq is not homogenous enough, and true signal is obscured by the background noise from other cell types that may have been

captured. I then continued my work to establish a homogeneous DA neuron population for ChIP-seq analysis.

To reach this end, we obtained a transgenic mouse line driving EGFP expression under the control of tyrosine hydroxylase. I optimized a protocol to dissect and dissociate neurons from the ventral midbrain of E15.5 embryos for FAC sorting so we could collect these EGFP neurons *ex vivo*. We showed that this population expresses DA markers in high levels and is therefore a good substrate for DA neuron ChIP-seq. However, these cells are present in very low numbers making up only about 2% of the total ventral midbrain neurons so further optimization of a small cell number ChIp-seq protocol was required. I finally completed ChIP-seq for H3K4me1 and H3K27ac on about 250,000 cells from these *ex vivo* sorted neurons to identify a catalog of putative DA enhancers. These putative enhancer show enrichment for conservation, are located near genes that are expressed in the *substantia nigra* and CNS as a whole, and are often located near genes associated with development of Parkinson's disease. This catalog must be further validated, but offers a step in the direction of furthering our understanding of transcriptional regulation in DA neurons.

6.2 Associating putative enhancers with their cognate genes

Throughout each of these chapters we have made the assumption that the putative enhancers identified will act to regulate the gene that is nearest to them. However, this assumption is not always true and further work should be done in each of these cases to determine which gene(s) each enhancer physically interacts with. Chromatin Conformation Capture (3C) has advanced significantly in recent years, allowing one to examine physical interactions taking place between distal regulatory elements and promoters (de Laat and Dekker, 2012). Characterizing the physical interactions underlying regulatory networks

would drastically improve our understanding of transcriptional regulation as whole, and the sequences necessary for the development and maintenance of our neuronal populations of interest.

For example, in the LMX1A and B study we delimited the non-coding search space by the genes flanking each LMX1 locus. Although this was the best method available at the time, it does not ensure that the non-coding sequences we studied are truly LMX1 enhancers. We hope to enrich for sequences that drive LMX1 expression by including the endogenous expression patterns of zebrafish homologs in our analysis and looking for overlapping patterns of expression. However, this approach is not particularly good for filtering out enhancers of genes that may have similar expression patterns. Specifically, the genes flanking LMX1A (PBX1 and RXRG) are also expressed in the CNS, thus reducing the certainty that the enhancers we identify truly target the LMX1A gene. 3C analysis would aid in identifying the true target of each enhancer and lead to a better understanding of the regulatory networks.

Similarly, when identifying enhancers genome-wide as in chapters 3-5 it is helpful to examine global expression patterns of the cells or tissues of interest (as in Gorkin et al., 2012). We were able to examine the correlation between putative enhancers and expressed genes in the analysis of human *substantia nigra*, however RNA-seq or microarray data is not currently available for the human hindbrain or mouse VM. We plan to complete RNA-seq in the *ex vivo* sorted population from mouse VM. Our hypothesis is that active enhancers are more likely to be located near genes that are highly expressed in our cell type. This analysis may also aid in correlating enhancers with their cognate genes by removing genes with little to no expression in the appropriate cell type from the list of possible regulatory targets.

6.3 Functional characterization of necessity and sufficiency of enhancers

Once we are able to determine precise transcriptional networks of enhancers and cognate promoters it becomes possible to start examining the necessity and sufficiency of a particular enhancer to the expression of its cognate gene. CRISPR-Cas9 technology is making it possible to edit the eukaryotic genomes easily and quickly (Mali et al., 2013). We can co-opt this technology in cell culture or *in vivo* to remove or modify specific enhancer sequences and characterize the changes in expression that result from these changes. This analysis would allow us to identify which regulatory sequences are required for the development and maintenance of a specific cell type and which are redundant. This information could then be combined with GWAS data to determine which sequences may be involved in the common variation that is responsible for human phenotypes (Welter et al., 2014).

Increasing functional characterization of enhancers is necessary, particularly for those active in disease relevant cell populations, such as DA neurons. These studies may be difficult to interpret if physical interaction maps are not available, however expression profiling will aid in the identification of gene targets for the subset of enhancers that are required. Identifying the appropriate cellular substrate for these studies may also be difficult, as it has been for finding a suitable DA neuron population for ChIP-seq, however ES and iPS cellular differentiation is constantly improving and characterization of neuronal subtypes has improved drastically (Frilling et al., 2010, Holmberg and Perlmann, 2012; Mong et al., 2014). Additionally, large cell numbers are not necessary for expression and viability assays as they once were for ChIP-seq, making these technically simpler experiments.

6.4 Translation of enhancer activity to human disease

The extreme end goal of my work has been to better understand the role sequence variation in driving common neurological diseases so that we may eventually identify therapies that can help protect and rehabilitate individuals with these disorders. We still have a long way to go for complete understanding, but I believe that the scientific community is making great strides towards these goals. During my thesis work hundreds of thousands of putative enhancers have been identified using ChIP-seq for 119 different proteins in 72 cell types and Dnase-seq in 125 cell types (The ENCODE Project Consortium, 2012). This massive amount of publically available data, along with smaller studies like mine, has proven that we are now capable of identifying catalogs of enhancers for cell types that are readily available. We can continue to identify more and more enhancer catalogs in specific cell types in this same way. However it is also important that future studies also focus on the functional aspects of identified enhancers, the impact of sequence variation within them and their role in human disease. Therefore, our efforts to define and catalog enhancers are not an end in themselves.

Despite this, my work in identifying neuronally expressed enhancers has contributed to an improved understanding of regulatory vocabularies and lays a framework for future studies. Although my early efforts to characterize Hb vocabularies may now appear rudimentary, they were among the first to comprehensively define regulatory encryption in specific tissues. They also pointed the way to my subsequenct efforts to identify cell-type dependent (DA neuron) regulatory vocabularies. These and related efforts continue to present technical challenges, but those too will be overcome. The results will be to illuminate the sequence basis of cell-dependent transcriptional control and facilitate prediction of the impact and phenotypic consequences of variation in regulatory noncoding sequences.

REFERENCES

Adams KA, Maida JM, Golden JA, Riddle RD. The transcription factor Lmx1b maintains Wnt1 expression within the isthmic organizer. Development. 2000 May;127(9):1857-67.

Addis RC, Hsu FC, Wright RL, Dichter MA, Coulter DA, Gearhart JD. Efficient conversion of astrocytes to functional midbrain dopaminergic neurons using a single polycistronic vector. PLoS One. 2011;6(12):e28719.

Adli M, Zhu J, Bernstein BE. Genome-wide chromatin maps derived from limited numbers of hematopoietic progenitors. Nat Methods. 2010 Aug;7(8):615-8.

Adli M, Bernstein BE. Whole-genome chromatin profiling from limited numbers of cells using nano-ChIP-seq. Nat Protoc. 2011 Sep 29;6(10):1656-68. Ahn KJ, Passero F Jr, Crenshaw EB 3rd. Otic mesenchyme expression of Cre recombinase directed by the inner ear enhancer of the Brn4/Pou3f4 gene. Genesis. 2009 Mar;47(3):137-41. doi: 10.1002/dvg.20454.

Aizawa H, Amo R, Okamoto H. Phylogeny and ontogeny of the habenular structure. Front Neurosci. 2011 Dec 21;5:138.

Amberger J, Bocchini CA, Scott AF, Hamosh A. McKusick's Online Mendelian Inheritance in Man (OMIM). Nucleic Acids Res. 2009 Jan;37 (Database issue): D793-6.

Amores A, Force A, Yan YL, Joly L, Amemiya C, Fritz A, Ho RK, Langeland J, Prince V, Wang YL, Westerfield M, Ekker M, Postlethwait JH. Zebrafish hox clusters and vertebrate genome evolution. Science. 1998 Nov 27;282(5394):1711-4.

Andreasen NC, Pierson R. The role of the cerebellum in schizophrenia. Biol Psychiatry. 2008 Jul 15;64(2):81-8.

Antonellis A, Huynh JL, Lee-Lin SQ, Vinton RM, Renaud G, Loftus SK, Elliot G, Wolfsberg TG, Green ED, McCallion AS, Pavan WJ. Identification of neural crest and glial enhancers at the mouse Sox10 locus through transgenesis in zebrafish. PLoS Genet. 2008 Sep 5;4(9):e1000174.

Aston-Jones G. Brain structures and receptors involved in alertness. Sleep Med. 2005 Jun;6 Suppl 1:S3-7.

Avila RL, Tevlin BR, Lees JP, Inouye H, Kirschner DA. Myelin structure and composition in zebrafish. Neurochem Res. 2007;32:197–209.

Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol. 1994;2:28-36.

Bailey TL, Gribskov M. Methods and statistics for combining motif match scores. J Comput Biol. 1998 Summer;5(2):211-21.

Banerji J, Rusconi S, Schaffner W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. Cell. 1981 Dec;27(2 Pt 1):299-308.

Barbazuk WB, Korf I, Kadavi C, Heyen J, Tate S, Wun E, Bedell JA, McPherson JD, Johnson SL. The syntenic relationship of the zebrafish and human genomes. Genome Res. 2000 Sep;10(9):1351-8.

Bardet PL, Schubert M, Horard B, Holland LZ, Laudet V, Holland ND, Vanacker JM. Expression of estrogen-receptor related receptors in amphioxus and zebrafish: implications for the evolution of posterior brain segmentation at the invertebrate-to-vertebrate transition. Evol Dev. 2005 May-Jun;7(3):223-33.

Barski A, Zhao K. Genomic location analysis by ChIP-Seq. J Cell Biochem. 2009 May 1;107(1):11-8.

Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol 1995 57: 289–300.

Benko S, Fantes JA, Amiel J, Kleinjan DJ, Thomas S, Ramsay J, Jamshidi N, Essafi A, Heaney S, Gordon CT, McBride D, Golzio C, Fisher M, Perry P, Abadie V, Ayuso C, Holder-Espinasse M, Kilpatrick N, Lees MM, Picard A, Temple IK, Thomas P, Vazquez MP, Vekemans M, Roest Crollius H, Hastie ND, Munnich A, Etchevers HC, Pelet A, Farlie PG, Fitzpatrick DR, Lyonnet S. Highly conserved non-coding elements on either side of SOX9 associated with Pierre Robin sequence. Nat Genet. 2009 Mar;41(3):359-64.

Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, Farnham PJ, Hirst M, Lander ES, Mikkelsen TS, Thomson JA. The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol. 2010 Oct;28(10):1045-8.

Berquin PC, Giedd JN, Jacobsen LK, Hamburger SD, Krain AL, Rapoport JL, Castellanos FX. Cerebellum in attention-deficit hyperactivity disorder: a morphometric MRI study. Neurology. 1998 Apr;50(4):1087-93.

Bhatia S, Bengani H, Fish M, Brown A, Divizia MT, de Marco R, Damante G, Grainger R, van Heyningen V, Kleinjan DA. Disruption of autoregulatory feedback by a mutation in a remote, ultraconserved PAX6 enhancer causes aniridia. Am J Hum Genet. 2013 Dec 5;93(6):1126-34.

Biedler JL, Helson L, Spengler BA. Morphology and growth, tumorigenicity, and cytogenetics of human neuroblastoma cells in continuous culture. Cancer Res. 1973 Nov; 33(11):2643-52.

Binot AC, Manfroid I, Flasse L, Winandy M, Motte P, Martial JA, Peers B, Voz ML. Nkx6.1 and nkx6.2 regulate alpha- and beta-cell formation in zebrafish by acting on pancreatic endocrine progenitor cells. Dev Biol. 2010 Apr 15;340(2):397-407.

Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J. Galaxy: a web-based genome analysis tool for experimentalists. Curr Protoc Mol Biol. 2010 Jan;Chapter 19:Unit 19.10.1-21.

Blow MJ, McCulley DJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, Afzal V, Bristow J, Ren B, Black BL, Rubin EM, Visel A, Pennacchio LA. ChIP-Seq identification of weakly conserved heart enhancers. Nat Genet. 2010 Sep;42(9): 806-10.

Bonn S, Zinzen RP, Perez-Gonzalez A, Riddell A, Gavin AC, Furlong EE. Cell type-specific chromatin immunoprecipitation from multicellular complex samples using BiTS-ChIP. Nat Protoc. 2012 Apr 26;7(5):978-94.

Bryne JC, Valen E, Tang MH, Marstrand T, Winther O, da Piedade I, Krogh A, Lenhard B, Sandelin A. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic Acids Res. 2008 Jan; 36(Database issue):D102-6.

Budick SA, O'Malley DM. Locomotor repertoire of the larval zebrafish: swimming, turning and prey capture. J Exp Biol. 2000 Sep;203(Pt 17):2565-79.

Bulger M, Groudine M. Functional and mechanistic diversity of distal transcription enhancers. Cell. 2011 Feb 4;144(3):327-39.

Burke TW, Kadonaga JT. Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. Genes Dev 1996;10:711–724.

Burke TW, Kadonaga JT. The downstream core promoter element, DPE, is conserved from Drosophila to humans and is recognized by TAFII60 of Drosophila. Genes Dev 1997; 11:3020–3031.

Buratowski S, Hahn S, Guarente L, Sharp PA. Five intermediate complexes in transcription initiation by RNA polymerase II. Cell. 1989 Feb 24;56(4):549-61.

Burzynski GM, Reed X, Taher L, Stine ZE, Matsui T, Ovcharenko I, McCallion AS. Systematic elucidation and in vivo validation of sequences enriched in hindbrain transcriptional control. Genome Res. 2012 Nov;22(11):2278-89.

Burzynski GM, Reed X, Maragh S, Matsui T, McCallion AS. Integration of genomic and functional approaches reveals enhancers at LMX1A and LMX1B. Mol Genet Genomics. 2013 Nov;288(11):579-89.

Caiazzo M, Dell'Anno MT, Dvoretskova E, Lazarevic D, Taverna S, Leo D, Sotnikova TD, Menegon A, Roncaglia P, Colciago G, Russo G, Carninci P, Pezzoli G, Gainetdinov RR, Gustincich S, Dityatev A, Broccoli V. Direct generation of functional dopaminergic neurons from mouse and human fibroblasts. Nature. 2011 Jul 3; 476(7359):224-7.

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS,Engström PG, Frith MC, Forrest AR, Alkema WB, Tan SL, Plessy C, Kodzius R, Ravasi T, KasukawaT, Fukuda S, Kanamori-Katayama M, Kitazume Y, Kawaji H, Kai C, Nakamura M, Konno H, NakanoK, Mottagui-Tabar S, Arner P, Chesi A, Gustincich S, Persichetti F, Suzuki H, Grimmond SM, WellsCA, Orlando V, Wahlestedt C, Liu ET, Harbers M, Kawai J, Bajic VB, Hume DA, Hayashizaki Y. Genome-wide analysis of mammalian promoter architecture and evolution. Nat. Genet 2006;38:626–635.

Cepeda-Nieto AC, Pfaff SL, Varela-Echavarría A. Homeodomain transcription factors in the development of subsets of hindbrain reticulospinal neurons. Mol Cell Neurosci. 2005 Jan;28(1):30-41.

Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM Trans Intell Syst Technol 2011 2: article 27.

Cheng CW, Yan CH, Choy SW, Hui MN, Hui CC, Cheng SH. Zebrafish homologue irx1a is required for the differentiation of serotonergic neurons. Dev Dyn. 2007 Sep;236(9): 2661-7.

Chepelev I, Wei G, Wangsa D, Tang Q, Zhao K. Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. Cell Res. 2012 Mar; 22(3):490- 503.

Chizhikov VV, Lindgren AG, Mishima Y, Roberts RW, Aldinger KA, Miesegaes GR, Currle DS, Monuki ES, Millen KJ. Lmx1a regulates fates and location of cells originating from the cerebellar rhombic lip and telencephalic cortical hem. Proc Natl Acad Sci USA. 2010 Jun 8;107(23):10725-30.

Cooper GM, Stone EA, Asimenos G; NISC Comparative Sequencing Program, Green ED, Batzoglou S, Sidow A. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 2005 Jul;15(7):901-13.

Cooper GM, Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet. 2011 Aug 18;12(9):628-40.

Cooper SJ, Trinklein ND, Anton ED, Nguyen L, Myers RM. Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. Genome Res 2006;16:1–10.

Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, Boyer LA, Young RA, Jaenisch R. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci USA. 2010 Dec 14;107(50):21931-6.

Dai JX, Johnson RL, Ding YQ. Manifold functions of the Nail-Patella Syndrome gene Lmx1b in vertebrate development. Dev Growth Differ. 2009 Apr;51(3):241-50.

de Kok YJ, Merkx GF, van der Maarel SM, Huber I, Malcolm S, Ropers HH, Cremers FP. A duplication/paracentric inversion associated with familial X-linked deafness (DFN3) suggests the presence of a regulatory element more than 400 kb upstream of the POU3F4 gene. Hum Mol Genet. 1995 Nov;4(11):2145-50.

de Laat W, Dekker J. 3C-based technologies to study the shape of the genome. Methods. 2012 Nov;58(3):189-91.

Ding YQ, Marklund U, Yuan W, Yin J, Wegman L, Ericson J, Deneris E, Johnson RL, Chen ZF. Lmx1b is essential for the development of serotonergic neurons. Nat Neurosci. 2003 Sep;6(9):933-8.

Doulazmi M, Frédéric F, Capone F, Becker-André M, Delhaye-Bouchaud N, Mariani J. A comparative study of Purkinje cells in two RORalpha gene mutant mice: staggerer and RORalpha(-/-).Brain Res Dev Brain Res. 2001 Apr 30;127(2):165-74.

Eckner R, Arany Z, Ewen M, Sellers W, Livingston DM. The adenovirus E1A-associated 300-kD protein exhibits properties of a transcriptional coactivator and belongs to an evolutionarily conserved family. Cold Spring Harb Symp Quant Biol. 1994; 59: 85-95.

Elia J, Gai X, Xie HM, Perin JC, Geiger E, Glessner JT, D'arcy M, deBerardinis R, Frackelton E, Kim C, Lantieri F, Muganga BM, Wang L, Takeda T, Rappaport EF,

Grant SF, Berrettini W, Devoto M, Shaikh TH, Hakonarson H, White PS. Rare structural variants found in attention-deficit hyperactivity disorder are preferentially associated with neurodevelopmental genes. Mol Psychiatry. 2010 Jun;15(6):637-46.

Elsen GE, Choi LY, Millen KJ, Grinblat Y, Prince VE. Zic1 and Zic4 regulate zebrafish roof plate specification and hindbrain ventricle morphogenesis. Dev Biol. 2008 Feb 15;314(2):376-92.

Emison ES, McCallion AS, Kashuk CS, Bush RT, Grice E, Lin S, Portnoy ME, Cutler DJ, Green ED, Chakravarti A. A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk. Nature. 2005 Apr 14;434 (7035):857-63.

ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007 Jun14;447(7146):799-816.

ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol. 2011 Apr;9(4):e1001046.

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, Ku M, Durham T, Kellis M, Bernstein BE. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011 May 5;473(7345):43-9.

Failli V, Bachy I, Rétaux S. Expression of the LIM-homeodomain gene Lmx1a (dreher) during development of the mouse nervous system. Mech Dev. 2002 Oct;118(1-2): 225-8.

Fantes J, Redeker B, Breen M, Boyle S, Brown J, Fletcher J, Jones S, Bickmore W, Fukushima Y, Mannens M, et al. Aniridia-associated cytogenetic rearrangements suggest that a position effect may cause the mutant phenotype. Hum Mol Genet. 1995 Mar;4(3): 415-22.

Ferri AL, Cavallaro M, Braida D, Di Cristofano A, Canta A, Vezzani A, Ottolenghi S, Pandolfi PP, Sala M, DeBiasi S, Nicolis SK. Sox2 deficiency causes neurodegeneration and impaired neurogenesis in the adult mouse brain. Development. 2004 Aug;131(15): 3805-19.

Ferri AL, Lin W, Mavromatakis YE, Wang JC, Sasaki H, Whitsett JA, Ang SL. Foxa1 and Foxa2 regulate multiple phases of midbrain dopaminergic neuron development in a dosage-dependent manner. Development. 2007 Aug;134(15): 2761-9.

Filippi A, Mahler J, Schweitzer J, Driever W. Expression of the paralogous tyrosine hydroxylase encoding genes th1 and th2 reveals the full complement of dopaminergic and noradrenergic neurons in zebrafish larval and juvenile brain. J Comp Neurol. 2010 Feb 15;518(4):423-38.

Filippi A, Mueller T, Driever W. vglut2 and gad expression reveal distinct patterns of dual GABAergic versus glutamatergic cotransmitter phenotypes of dopaminergic and noradrenergic neurons in the zebrafish brain. J Comp Neurol. 2014 Jun 15; 522(9):2019-37.

Fisher S, Grice EA, Vinton RM, Bessling SL, McCallion AS. Conservation of RET regulatory function from human to zebrafish without sequence similarity. Science. 2006 Apr 14;312(5771):276-9.

Fisher S, Grice EA, Vinton RM, Bessling SL, Urasaki A, Kawakami K, McCallion AS. Evaluating the biological relevance of putative enhancers using Tol2 transposon-mediated transgenesis in zebrafish. Nat Protoc. 2006;1(3):1297-305.

FitzGerald PC, Sturgill D, Shyakhtenko A, Oliver B, Vinson C. Comparative genomics of Drosophila and human core promoters. Genome Biol 2006;7:R53.

Fletez-Brant C, Lee D, McCallion AS, Beer MA. kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W544-56. doi: 10.1093/nar/gkt519.

Friling S, Andersson E, Thompson LH, Jönsson ME, Hebsgaard JB, Nanou E, Alekseenko Z, Marklund U, Kjellander S, Volakakis N, Hovatta O, El Manira A, Björklund A, Perlmann T, Ericson J. Efficient production of mesencephalic dopamine neurons by Lmx1a expression in embryonic stem cells. Proc Natl Acad Sci USA. 2009 May 5;106 (18):7613-8.

Gadd MS, Bhati M, Jeffries CM, Langley DB, Trewhella J, Guss JM, Matthews JM. Structural basis for partial redundancy in a class of transcription factors, the LIM homeodomain proteins, in neural cell type specification. J Biol Chem. 2011 Dec 16;286(50):42971-80.

Gary DS, Mattson MP. Integrin signaling via the PI3-kinase-Akt pathway increases neuronal resistance to glutamate-induced apoptosis. J Neurochem. 2001 Mar;76(5): 1485-96.

Gary DS, Milhavet O, Camandola S, Mattson MP. Essential role for integrin linked kinase in Akt-mediated integrin survival signaling in hippocampal neurons. J Neurochem. 2003 Feb;84(4):878-90.

Gary DS, Malone M, Capestany P, Houdayer T, McDonald JW. Electrical stimulation promotes the survival of oligodendrocytes in mixed cortical cultures. J Neurosci Res. 2012 Jan;90(1):72-83.

Gates MA, Kim L, Egan ES, Cardozo T, Sirotkin HI, Dougan ST, Lashkari D, Abagyan R, Schier AF, Talbot WS. A genetic linkage map for zebrafish: comparative analysis and localization of genes and expressed sequences. Genome Res. 1999 Apr;9(4):334-47.

Gershenzon NI, Trifonov EN, Ioshikhes IP. The features of Drosophila core promoters revealed by statistical analysis. BMC Genomics 2006;7:161.

Ghysen A. The origin and evolution of the nervous system. Int J Dev Biol. 2003;47(7-8):555-62.

Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE et al. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature. 2004 Apr 1;428(6982):493-521.

Giguère V. To ERR in the estrogen pathway. Trends Endocrinol Metab. 2002 Jul;3(5):220-225.

Gilfillan GD, Hughes T, Sheng Y, Hjorthaug HS, Straub T, Gervin K, Harris JR, Undlien DE, Lyle R. Limitations and possibilities of low cell number ChIP-seq. BMC Genomics. 2012 Nov 21;13:645.

Goecks J, Nekrutenko A, Taylor J; Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 2010;11(8):R86.

Goldberg, ML. Ph.D. Thesis. Stanford University; Stanford, CA, U.S.A.: 1979.

Gompel N, Prud'homme B, Wittkopp PJ, Kassner VA, Carroll SB. Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila. Nature. 2005 Feb 3;433(7025):481-7.

Gong S, Zheng C, Doughty ML, Losos K, Didkovsky N, Schambra UB, Nowak NJ, Joyner A, Leblanc G, Hatten ME, Heintz N. A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. Nature. 2003 Oct 30; 425(6961): 917-25.

Gorkin DU, Lee D, Reed X, Fletez-Brant C, Bessling SL, Loftus SK, Beer MA, Pavan WJ, McCallion AS. Integration of ChIP-seq and machine learning reveals enhancers and a predictive regulatory sequence vocabulary in melanocytes. Genome Res. 2012 Nov;22(11):2290-301.

Gorkin, DU. Ph.D. Thesis. Johns Hopkins University; Baltimore, MD, U.S.A.: 2013.

Graham V, Khudyakov J, Ellis P, Pevny L. SOX2 functions to maintain neural progenitor identity. Neuron. 2003 Aug 28;39(5):749-65.

Grice EA, Rochelle ES, Green ED, Chakravarti A, McCallion AS. Evaluation of the RET regulatory landscape reveals the biological relevance of a HSCR-implicated enhancer. Hum Mol Genet. 2005 Dec 15;14(24):3837-45.

Guo C, Qiu HY, Huang Y, Chen H, Yang RQ, Chen SD, Johnson RL, Chen ZF, Ding YQ. Lmx1b is essential for Fgf8 and Wnt1 expression in the isthmic organizer during tectum and cerebellum development in mice. Development. 2007 Jan;134(2):317-25.

Guo S. Linking genes to brain, behavior and neurological diseases: what can we learn from zebrafish? Genes Brain Behav. 2004 Apr;3(2):63-74.

Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn 2002 46: 389–422.

Hawkins RD, Hon GC, Yang C, Antosiewicz-Bourget JE, Lee LK, Ngo QM, Klugman S, Ching KA, Edsall LE, Ye Z, Kuan S, Yu P, Liu H, Zhang X, Green RD, Lobanenkov VV, Stewart R, Thomson JA, Ren B. Dynamic chromatin states in human ES cells reveal potential regulatory sequences and genes involved in pluripotency. Cell Res. 2011 Oct;21(10):1393-409.

He A, Kong SW, Ma Q, Pu WT. Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart. Proc Natl Acad Sci USA. 2011 Apr 5;108(14):5632-7.

Heintz N. Gene expression nervous system atlas (GENSAT). Nat Neurosci. 2004 May;7(5): 483.

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE, Ren B. Distinct

and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet. 2007 Mar;39(3):311-8.

Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, Ching KA, Antosiewicz-Bourget JE, Liu H, Zhang X, Green RD, Lobanenkov V, Stewart R, Thomson JA, Crawford GE, Kellis M, Ren B. Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature. 2009 May 7; 459(7243):108-12.

Hitzemann R, Qian Y, Hitzemann B. Dopamine and acetylcholine cell density in the neuroleptic responsive (NR) and neuroleptic nonresponsive (NNR) lines of mice. J Pharmacol Exp Ther. 1993 Jul;266(1):431-8.

Hobert O, Westphal H. Functions of LIM-homeobox genes. Trends Genet. 2000 Feb;16(2): 75-83.

Holmberg J, Perlmann T. Maintaining differentiated cellular identity. Nat Rev Genet. 2012 May 18;13(6):429-39.

Holmqvist PH, Mannervik M. Genomic occupancy of the transcriptional co-activators p300 and CBP. Transcription. 2013 Jan-Feb;4(1):18-23.

Holzschuh J, Barrallo-Gimeno A, Ettl AK, Durr K, Knapik EW, Driever W. Noradrenergic neurons in the zebrafish hindbrain are induced by retinoic acid and require tfap2a for expression of the neurotransmitter phenotype. Development. 2003 Dec;130(23): 5741-54.

Horak CE, Mahajan MC, Luscombe NM, Gerstein M, Weissman SM, Snyder M. GATA-1 binding sites mapped in the beta-globin locus by using mammalian chIp-chip analysis. Proc Natl Acad Sci USA. 2002 Mar 5;99(5):2924-9.

Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D. The UCSC Known Genes. Bioinformatics. 2006 May 1;22(9):1036-46.

Huang M, Sage C, Li H, Xiang M, Heller S, Chen ZY. Diverse expression patterns of LIM-homeodomain transcription factors (LIM-HDs) in mammalian inner ear development. Dev Dyn. 2008 Nov;237(11):3305-12.

Huang Z, Dore LC, Li Z, Orkin SH, Feng G, Lin S, Crispino JD. GATA-2 reinforces megakaryocyte development in the absence of GATA-1. Mol Cell Biol. 2009 Sep;29(18):5168-80.

Huettner JE, Baughman RW. Primary culture of identified neurons from the visual cortex of postnatal rats. J Neurosci. 1986 Oct;6(10):3044-60.

Ibáñez-Sandoval O, Tecuapetla F, Unal B, Shah F, Koós T, Tepper JM. Electrophysiological and morphological characteristics and synaptic connectivity of tyrosine hydroxylase-expressing neurons in adult mouse striatum. J Neurosci. 2010 May 19;30(20):6999-7016.

Inoka S. Balasooriya and K. Wimalasena. Are SH-SY5Y and MN9D cell lines truly dopaminergic? FASEB J. April 2007. 21 (Meeting Abstract Supplement) A1274.

Ji H, Jiang H, Ma W, Johnson DS, Myers RM, Wong WH. An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nat Biotechnol. 2008 Nov;26(11):1293-300.

Juven-Gershon T1, Kadonaga JT. Regulation of gene expression via the core promoter and the basal transcriptional machinery. Dev Biol. 2010 Mar 15;339(2):225-9.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 2004 Jan1;32 (Database issue):D493-6.

Karolchik D, Kuhn RM, Baertsch R, Barber GP, Clawson H, Diekhans M, Giardine B, Harte RA, Hinrichs AS, Hsu F, Kober KM, Miller W, Pedersen JS, Pohl A, Raney BJ, Rhead B, Rosenbloom KR, Smith KE, Stanke M, Thakkapallayil A, Trumbower H, Wang T, Zweig AS, Haussler D, Kent WJ. The UCSC Genome Browser Database: 2008 update. Nucleic Acids Res. 2008 Jan;36(Database issue):D773-9.

Kawai H, Arata N, Nakayasu H. Three-dimensional distribution of astrocytes in zebrafish spinal cord. Glia. 2001;36:406–413.

Kelberman D, de Castro SC, Huang S, Crolla JA, Palmer R, Gregory JW, Taylor D, Cavallo L, Faienza MF, Fischetto R, Achermann JC, Martinez-Barbera JP, Rizzoti K, Lovell-Badge R, Robinson IC, Gerrelli D, Dattani MT. SOX2 plays a critical role in the pituitary, forebrain, and eye during human embryonic development. J Clin Endocrinol Metab. 2008 May;93(5):1865-73.

Khaliq ZM, Bean BP. Pacemaking in dopaminergic ventral tegmental area neurons: depolarizing drive from background and voltage-dependent sodium conductances. J Neurosci. 2010 May 26;30(21):7401-13

Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, Richmond TA, Wu Y, Green RD, Ren B. A high resolution map of active promoters in the human genome. Nature 2005;436:876– 880.

Kimmel CB, Ballard WW, Kimmel SR, Ullmann B, Schilling TF. Stages of embryonic development of the zebrafish. Dev Dyn. 1995 Jul;203(3):253-310.

Kittler J, Hatef M, Duin RPW, Matas J. On combining classifiers. IEEE Trans Pattern Anal Mach Intell 1998 20: 226–239.

Kiyota T, Kato A, Altmann CR, Kato Y. The POU homeobox protein Oct-1 regulates radial glia formation downstream of Notch signaling. Dev Biol. 2008 Mar15;315(2):579-92.

Kleinjan DJ1, Coutinho P. Cis-ruption mechanisms: disruption of cis-regulatory control as a cause of human genetic disease. Brief Funct Genomic Proteomic. 2009 Jul;8(4):317-32.

Koo SK, Hill JK, Hwang CH, Lin ZS, Millen KJ, Wu DK. Lmx1a maintains proper neurogenic, sensory, and non-sensory domains in the mammalian inner ear. Dev Biol. 2009 Sep 1;333(1):14-25.

Koulen P, Janowitz T, Johnston LD, Ehrlich BE. Conservation of localization patterns of IP (3) receptor type 1 in cerebellar Purkinje cells across vertebrate species. J Neurosci Res. 2000;61:493–499.

Kuwada JY. Development of the zebrafish nervous system: genetic analysis and manipulation. Curr Opin Neurobiol. 1995 Feb;5(1):50-4.

Lagrange T, Kapanidis AN, Tang H, Reinberg D, Ebright RH. New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. Genes Dev 1998;12:34–44.

Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, Chen Y, DeSalvo G, Epstein C, Fisher-Aylor KI, Euskirchen G, Gerstein M, Gertz J, Hartemink AJ, Hoffman MM, Iyer VR, Jung YL, Karmakar S, Kellis M, Kharchenko PV, Li Q, Liu T, Liu XS, Ma L, Milosavljevic A, Myers RM, Park PJ, Pazin MJ, Perry MD, Raha D, Reddy TE, Rozowsky J, Shoresh N, Sidow A, Slattery M, Stamatoyannopoulos JA, Tolstorukov MY, White KP, Xi S, Farnham PJ, Lieb JD, Wold BJ, Snyder M. ChIP-seq guidelines and practices of the ENCODE and mod ENCODE consortia. Genome Res. 2012 Sep;22(9):1813-31.

Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012, 9:357-359.

Lee D, Karchin R, Beer MA. Discriminative prediction of mammalian enhancers from DNA sequence. Genome Res. 2011 Dec;21(12):2167-80.

Lee HS, Bae EJ, Yi SH, Shim JW, Jo AY, Kang JS, Yoon EH, Rhee YH, Park CH, Koh HC, Kim HJ, Choi HS, Han JW, Lee YS, Kim J, Li JY, Brundin P, Lee SH. Foxa2 and Nurr1 synergistically yield A9 nigral dopamine neurons exhibiting improved differentiation, function, and cell survival. Stem Cells. 2010 Mar 31;28(3):501-12.

Leleu M, Lefebvre G, Rougemont J. Processing and analyzing ChIP-seq data: from short reads to regulatory interactions. Brief Funct Genomics. 2010 Dec;9(5-6):466-76.

Lettice LA, Horikoshi T, Heaney SJ, van Baren MJ, van der Linde HC, Breedveld GJ, Joosse M, Akarsu N, Oostra BA, Endo N, Shibata M, Suzuki M, Takahashi E, Shinka T, Nakahori Y, Ayusawa D, Nakabayashi K, Scherer SW, Heutink P, Hill RE, Noji S. Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. Proc Natl Acad Sci USA. 2002 May 28;99(11):7548-53.

Lettice LA, Heaney SJ, Purdie LA, Li L, de Beer P, Oostra BA, Goode D, Elgar G, Hill RE, de Graaff E. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum Mol Genet. 2003 Jul 15;12(14): 1725-35.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16): 2078-9.

Lim CY, Santoso B, Boulay T, Dong E, Ohler U, Kadonaga JT. The MTE, a new core promoter element for transcription by RNA polymerase II. Genes Dev 2004;18:1606-1617.

Lin MK, Farrer MJ. Genetics and genomics of Parkinson's disease. Genome Med. 2014 Jun 30; 6(6):48.

Lin W, Metzakopian E, Mavromatakis YE, Gao N, Balaskas N, Sasaki H, Briscoe J, Whitsett JA, Goulding M, Kaestner KH, Ang SL. Foxa1 and Foxa2 function both upstream of and cooperatively with Lmx1a and Lmx1b in a feed forward loop promoting

mesodiencephalic dopaminergic neuron development. Dev Biol. 2009 Sep 15; 333(2):386-96.

Liu T, Ortiz JA, Taing L, Meyer CA, Lee B, Zhang Y, Shin H, Wong SS, Ma J, Lei Y, Pape UJ, Poidinger M, Chen Y, Yeung K, Brown M, Turpaz Y, Liu XS. Cistrome: an integrative platform for transcriptional regulation studies. Genome Biol. 2011 Aug 22; 12(8):R83.

Liu YR, Laghari ZA, Novoa CA, Hughes J, Webster JR, Goodwin PE, Wheatley SP, Scotting PJ. Sox2 acts as a transcriptional repressor in neural stem cells. BMC Neurosci. 2014 Aug 8;15:95.

Mahony S, Benos PV. STAMP: a web tool for exploring DNA-binding motif similarities. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W253-8.

Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM. RNA-guided human genome engineering via Cas9. Science. 2013 Feb 15;339(6121):823-6.

Marei HE, Althani A, Afifi N, Michetti F, Pescatori M, Pallini R, Casalbore P, Cenciarelli C, Schwartz P, Ahmed AE. Gene expression profiling of embryonic human neural stem cells and dopaminergic neurons from adult human substantia nigra. PLoS One. 2011;6(12):e28420.

Matys V, Fricke E, Geffers R, Gössling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Münch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E. TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res. 2003 Jan 1;31(1):374-8.

Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. 2006 Jan 1; 34(Database issue):D108-10.

Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, Shafer A, Neri F, Lee K, Kutyavin T, Stehling-Sun S, Johnson AK, Canfield TK, Giste E, Diegel M, Bates D, Hansen RS, Neph S, Sabo PJ, Heimfeld S, Raubitschek A, Ziegler S, Cotsapas C, Sotoodehnia N, Glass I, Sunyaev SR, Kaul R, Stamatoyannopoulos JA. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012 Sep 7;337(6099):1190-5.

McGaughey DM, Stine ZE, Huynh JL, Vinton RM, McCallion AS. Asymmetrical distribution of non-conserved regulatory sequences at PHOX2B is reflected at the ENCODE loci and illuminates a possible genome-wide trend. BMC Genomics. 2009 Jan 7;10:8. McGaughey DM, Vinton RM, Huynh J, Al-Saif A, Beer MA, McCallion AS. Metrics of sequence constraint overlook regulatory sequences in an exhaustive analysis at phox2b. Genome Res. 2008 Feb;18(2):252-60.

McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol. 2010 May;28(5):495-501.

McMahon C, Gestri G, Wilson SW, Link BA. Lmx1b is essential for survival of periocular mesenchymal cells and influences Fgf-mediated retinal patterning in zebrafish. Dev Biol. 2009 Aug 15;332(2):287-98.

Mercer TR, Edwards SL, Clark MB, Neph SJ, Wang H, Stergachis AB, John S, Sandstrom R, Li G, Sandhu KS, Ruan Y, Nielsen LK, Mattick JS, Stamatoyannopoulos JA. DNaseI-hypersensitive exons colocalize with promoters and distal regulatory elements. Nat Genet. 2013 Aug;45(8):852-9.

Meyer-Lindenberg A, Miletich RS, Kohn PD, Esposito G, Carson RE, Quarantelli M, Weinberger DR, Berman KF. Reduced prefrontal activity predicts exaggerated striatal dopaminergic function in schizophrenia. Nat Neurosci. 2002 Mar;5(3):267-71.

Miguez A, Ducret S, Di Meglio T, Parras C, Hmidan H, Haton C, Sekizar S, Mannioui A, Vidal M, Kerever A, Nyabi O, Haigh J, Zalc B, Rijli FM, Thomas JL. Opposing roles for Hoxa2 and Hoxb2 in hindbrain oligodendrocyte patterning. J Neurosci. 2012 Nov 28; 32(48):17172-85.

Miller W, Makova KD, Nekrutenko A, Hardison RC. Comparative genomics. Annu Rev Genomics Hum Genet. 2004;5:15-56.

Millonig JH, Millen KJ, Hatten ME. The mouse Dreher gene Lmx1a controls formation of the roof plate in the vertebrate CNS. Nature. 2000 Feb 17;403(6771):764-9.

Mishima Y1, Lindgren AG, Chizhikov VV, Johnson RL, Millen KJ. Overlapping function of Lmx1a and Lmx1b in anterior hindbrain roof plate formation and cerebellar growth. J Neurosci. 2009 Sep 9;29(36):11377-84.

Mong J, Panman L, Alekseenko Z, Kee N, Stanton LW, Ericson J, Perlmann T. Transcription factor-induced lineage programming of noradrenaline and motor neurons from embryonic stem cells. Stem Cells. 2014 Mar;32(3):609-22.

Mueller T, Wullimann MF. Anatomy of neurogenesis in the early zebrafish brain. Brain Res Dev Brain Res. 2003 Jan 10;140(1):137-55.

Mueller T, Vernier P, Wullimann MF. A phylotypic stage in vertebrate brain development: GABA cell patterns in zebrafish compared with mouse. J Comp Neurol. 2006 Feb 1;494(4):620-34.

Nagatsu T, Levitt M, Udenfriend S. Tyrosine hydroxylase: the initial step in norepinephrine biosynthesis. J Biol Chem. 1964 Sep;239:2910-7.

Nakatani T, Kumai M, Mizuhara E, Minaki Y, Ono Y. Lmx1a and Lmx1b cooperate with Foxa2 to coordinate the specification of dopaminergic neurons and control of floor plate cell differentiation in the developing mesencephalon. Dev Biol. 2010 Mar 1;339(1): 101-13.

Naranjo S, Voesenek K, de la Calle-Mustienes E, Robert-Moreno A, Kokotas H, Grigoriadou M, Economides J, Van Camp G, Hilgert N, Moreno F, Alsina B, Petersen MB, Kremer H, Gómez-Skarmeta JL. Multiple enhancers located in a 1-Mb region upstream of POU3F4 promote expression during inner ear development and may be required for hearing. Hum Genet. 2010 Oct; 128(4):411-9.

Narlikar L, Gordân R, Hartemink AJ. A nucleosome-guided map of transcription factor binding sites in yeast. PLoS Comput Biol. 2007 Nov;3(11):e215.

Narlikar L, Sakabe NJ, Blanski AA, Arimura FE, Westlund JM, Nobrega MA, Ovcharenko I. Genome-wide discovery of human heart enhancers. Genome Res. 2010 Mar;20(3): 381-92.

Nelson SB, Janiesch C, Sander M. Expression of Nkx6 genes in the hindbrain and gut of the developing mouse. J Histochem Cytochem. 2005 Jun;53(6):787-90.

Nichols DH, Pauley S, Jahan I, Beisel KW, Millen KJ, Fritzsch B. Lmx1a is required for segregation of sensory epithelia and normal ear histogenesis and morphogenesis. Cell Tissue Res. 2008 Dec;334(3):339-58.

Nolis IK, McKay DJ, Mantouvalou E, Lomvardas S, Merika M, Thanos D. Transcription factors mediate long-range enhancer-promoter interactions. Proc Natl Acad Sci USA. 2009 Dec 1;106(48):20222-7.

Noonan JP, McCallion AS. Genomics of long-range regulatory elements. Annu Rev Genomics Hum Genet. 2010;11:1-23.

O'Hara FP, Beck E, Barr LK, Wong LL, Kessler DS, Riddle RD. Zebrafish Lmx1b.1 and Lmx1b.2 are required for maintenance of the isthmic organizer. Development. 2005 Jul;132(14):3163-73.

Ohler U, Liao GC, Niemann H, Rubin GM. Computational analysis of core promoters in the Drosophila genome. Genome Biol 2002;3:RESEARCH0087.

Ohnemus S, Bobola N, Kanzler B, Mallo M. Different levels of Hoxa2 are required for particular developmental processes. Mech Dev. 2001 Oct;108(1-2):135-47.

O'Neill LP, Turner BM. Histone H4 acetylation distinguishes coding regions of the human genome from heterochromatin in a differentiation-dependent but transcription-independent manner. EMBO J. 1995 Aug 15;14(16):3946-57.

O'Neill LP, Turner BM. Immunoprecipitation of chromatin. Methods Enzymol. 1996;274: 189-97.

Ong CT, Corces VG. Enhancer function: new insights into the regulation of tissue-specific gene expression. Nat Rev Genet. 2011 Apr;12(4):283-93.

Panne D, Maniatis T, Harrison SC. An atomic model of the interferon-beta enhanceosome. Cell. 2007 Jun 15;129(6):1111-23.

Pattyn A, Vallstedt A, Dias JM, Sander M, Ericson J. Complementary roles for Nkx6 and Nkx2 class proteins in the establishment of motoneuron identity in the hindbrain. Development. 2003 Sep;130(17):4149-59.

Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, Plajzer-Frick I, Akiyama J, De Val S, Afzal V, Black BL, Couronne O, Eisen MB, Visel A, Rubin EM. In vivo enhancer analysis of human conserved non-coding sequences. Nature. 2006 Nov 23;444(7118):499-502.

Peri F, Nusslein-Volhard C. Live imaging of neuronal degradation by microglia reveals a role for v0-ATPase a1 in phagosomal fusion in vivo. Cell. 2008;133:916–927.

Plaza S, Dozier C, Langlois MC, Saule S. Identification and characterization of a neuroretina-specific enhancer element in the quail Pax-6 (Pax-QNR) gene. Mol Cell Biol. 1995 Feb;15(2):892-903.

Prakash N, Wurst W. Development of dopaminergic neurons in the mammalian brain. Cell Mol Life Sci. 2006 Jan;63(2):187-206.

Pruitt KD, Tatusova T, Klimke W, Maglott DR. NCBI Reference Sequences: current status, policy and new initiatives. Nucleic Acids Res. 2009 Jan;37(Database issue):D32-6.

Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J. A unique chromatin signature uncovers early developmental enhancers in humans. Nature. 2011 Feb 10;470(7333):279-83.

Rick CE, Ebert A, Virag T, Bohn MC, Surmeier DJ. Differentiated dopaminergic MN9D cells only partially recapitulate the electrophysiological properties of midbrain dopaminergic neurons. Dev Neurosci. 2006;28(6):528-37.

Rink E, Wullimann MF. Connections of the ventral telencephalon and tyrosine hydroxylase distribution in the zebrafish brain (Danio rerio) lead to identification of an ascending dopaminergic system in a teleost. Brain Res Bull. 2002 Feb-Mar 1; 57(3-4):385-7.

Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, Thiessen N, Griffith OL, He A, Marra M, Snyder M, Jones S. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods. 2007 Aug; 4(8):651-7.

Robinson TE, Berridge KC. The neural basis of drug craving: an incentive-sensitization theory of addiction. Brain Res Brain Res Rev. 1993 Sep-Dec;18(3):247-91.

Ross RA, Spengler BA, Biedler JL. Coordinate morphological and biochemical interconversion of human neuroblastoma cells. J Natl Cancer Inst. 1983 Oct;71(4): 741-7.

Saucedo-Cardenas O, Quintana-Hau JD, Le WD, Smidt MP, Cox JJ, De Mayo F, Burbach JP, Conneely OM. Nurr1 is essential for the induction of the dopaminergic phenotype and the survival of ventral mesencephalic late dopaminergic precursor neurons. Proc Natl Acad Sci USA. 1998 Mar 31;95(7):4013-8.

Savitt JM, Dawson VL, Dawson TM. Diagnosis and treatment of Parkinson disease: molecules to medicine. J Clin Invest. 2006 Jul;116(7):1744-54.

Schneider-Maunoury S1, Gilardi-Hebenstreit P, Charnay P. How to build a vertebrate hindbrain. Lessons from genetics. C R Acad Sci III. 1998 Oct;321(10):819-34.

Serinagaoglu Y, Zhang R, Zhang Y, Zhang L, Hartt G, Young AP, Oberdick J. A promoter element with enhancer properties, and the orphan nuclear receptor RORalpha, are required for Purkinje cell-specific expression of a Gi/o modulator. Mol Cell Neurosci. 2007 Mar;34(3):324-42.

Serra HG, Duvick L, Zu T, Carlson K, Stevens S, Jorgensen N, Lysholm A, Burright E, Zoghbi HY, Clark HB, Andresen JM, Orr HT. RORalpha-mediated Purkinje cell

development determines disease severity in adult SCA1 mice. Cell. 2006 Nov 17;127(4):697-708.

Shaw MW, Falls HF, Neel JV. Congenital Aniridia. Am J Hum Genet. 1960 Dec;12(4Pt1):389- 415.

Shawe-Taylor J, Cristianini N. On the generalisation of soft margin algorithms. IEEE Trans Inf Theory 2002 48: 2721–2735.

Sheffield NC, Furey TS. Identifying and characterizing regulatory sequences in the human genome with chromatin accessibility assays. Genes (Basel). 2012 Oct15;3(4):651-70.

Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV, Ren B. A map of the cis-regulatory sequences in the mouse genome. Nature. 2012 Aug 2;488(7409):116-20.

Shirasaki R, Pfaff SL. Transcriptional codes and the control of neuronal identity. Annu Rev Neurosci. 2002;25:251-81.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005 Aug;15(8):1034-50.

Singleton AB, Farrer MJ, Bonifati V. The genetics of Parkinson's disease: progress and therapeutic implications. Mov Disord. 2013 Jan;28(1):14.

Smale ST, Baltimore D. The “initiator” as a transcription control element. Cell. 1989;57:103–113.

Smits SM, Ponnio T, Conneely OM, Burbach JP, Smidt MP. Involvement of Nurr1 in specifying the neurotransmitter identity of ventral midbrain dopaminergic neurons. Eur J Neurosci. 2003 Oct;18(7):1731- 8.

Song L, Crawford GE. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harb Protoc. 2010 Feb;2010(2):pdb.prot5384.

Spaniol P, Bornmann C, Hauptmann G, Gerster T. Class III POU genes of zebrafish are predominantly expressed in the central nervous system. Nucleic Acids Res. 1996 Dec 15;24(24):4874-81.

Stott SR, Metzakopian E, Lin W, Kaestner KH, Hen R, Ang SL. Foxa1 and foxa2 are required for the maintenance of dopaminergic properties in ventral midbrain neurons at late embryonic stages. J Neurosci. 2013 May 1;33(18):8022-34.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005 Oct 25;102(43):15545-50.

Tennyson VM, Gershon MD, Wade PR, Crotty DA, Wolgemuth DJ. Fetal development of the enteric nervous system of transgenic mice that overexpress the Hoxa-4 gene. Dev Dyn. 1998 Mar;211(3):269-91.

Thisse C, Thisse B, Schilling TF, Postlethwait JH.Structure of the zebrafish snail1 gene and its expression in wild-type, spadetail and no tail mutant embryos. Development. 1993 Dec;119(4):1203-15.

Thisse C, Degrave A, Kryukov GV, Gladyshev VN, Obrecht-Pflumio S, Krol A, Thisse B, Lescure A. Spatial and temporal expression patterns of selenoprotein genes during embryogenesis in zebrafish. Gene Expr Patterns. 2003 Aug;3(4):525-32.

Thisse B, Heyer V, Lux A, Alunni V, Degrave A, Seiliez I, Kirchner J, Parkhill JP, Thisse C. Spatial and temporal expression of the zebrafish genome by large-scale in situ hybridization screening. Methods Cell Biol. 2004;77:505-19.

Thisse B, Thisse C. Fast release clones: A high throughput expression analysis. In ZFIN Direct Data Submission (http://zfin.org) 2004.

Thomas MC, Chiang CM. The general transcription machinery and general cofactors. Crit Rev Biochem Mol Biol. 2006 May-Jun;41(3):105-78.

Thompson JA, Ziman M. Pax genes during neural development and their potential role in neuroregeneration. Prog Neurobiol. 2011 Nov;95(3):334-51.

Trinklein ND, Aldred SF, Hartman SJ, Schroeder DI, Otillar RP, Myers RM. An abundance of bidirectional promoters in the human genome. Genome Res. 2004 Jan;14(1):62-6.

Tsang AP, Visvader JE, Turner CA, Fujiwara Y, Yu C, Weiss MJ, Crossley M, Orkin SH. FOG, a multitype zinc finger protein, acts as a cofactor for transcription factor GATA-1 in erythroid and megakaryocytic differentiation. Cell. 1997 Jul 11; 90(1):109-19.

Tsompana M, Buck MJ. Chromatin accessibility: a window into the genome. Epigenetics Chromatin. 2014 Nov 20;7(1):33. doi: 10.1186/1756-8935-7-33.

Tzschentke TM, Schmidt WJ. Functional relationship among medial prefrontal cortex, nucleus accumbens, and ventral tegmental area in locomotion and reward. Crit Rev Neurobiol. 2000;14(2):131-42. Review.

Vanacker JM, Pettersson K, Gustafsson JA, Laudet V. Transcriptional targets shared by estrogen receptor- related receptors (ERRs) and estrogen receptor (ER) alpha, but not by ERbeta. EMBO J. 1999 Aug 2;18(15):4270-9.

Vanden Berghe W, De Bosscher K, Boone E, Plaisance S, Haegeman G. The nuclear factor-kappaB engages CBP/p300 and histone acetyltransferase activity for transcriptional activation of the interleukin-6 gene promoter. J Biol Chem. 1999 Nov 5;274(45): 32091-8.

Verrijzer CP, Van der Vliet PC. POU domain transcription factors. Biochim Biophys Acta. 1993 Apr 29;1173(1):1-21.

Visel A, Minovitsky S, Dubchak I, Pennacchio LA. VISTA Enhancer Browser--a database of tissue-specific human enhancers. Nucleic Acids Res. 2007 Jan;35(Database issue): D88-92.

Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, Afzal V, Ren B, Rubin EM, Pennacchio LA. ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature. 2009 Feb 2;457 (7231): 854-8.

Visel A, Rubin EM, Pennacchio LA. Genomic views of distant-acting enhancers. Nature. 2009 Sep 10;461(7261):199-205.

Vollrath D, Jaramillo-Babb VL, Clough MV, McIntosh I, Scott KM, Lichter PR, Richards JE. Loss-of-function mutations in the LIM-homeodomain gene, LMX1B, in nail-patella syndrome. Hum Mol Genet. 1998 Jul;7(7):1091-8.

Wang W, Zhong J, Wang YQ. Comparative genomic analysis reveals the evolutionary conservation of Pax gene family. Genes Genet Syst. 2010;85(3):193-206.

Waskiewicz AJ, Rikhof HA, Hernandez RE, Moens CB. Zebrafish Meis functions to stabilize Pbx proteins and regulate hindbrain patterning. Development. 2001 Nov; 128(21): 4139-51.

Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, Parkinson H. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014 Jan;42 (Database issue): D1001-6.

Wen L, Wei W, Gu W, Huang P, Ren X, Zhang Z, Zhu Z, Lin S, Zhang B. Visualization of monoaminergic neurons and neurotoxicity of MPTP in live transgenic zebrafish. Dev Biol. 2008 Feb 1;314(1):84-92.

Westerfield M, McMurray JV, Eisen JS. Identified motoneurons and their innervation of axial muscles in the zebrafish. J Neurosci. 1986;6:2267–2277.

Westerfield M. The zebrafish book: a guide for the laboratory use of zebrafish (Danio rerio). University of Oregon Press, Eugene. 2000.

Yan CH, Levesque M, Claxton S, Johnson RL, Ang SL. Lmx1a and lmx1b function cooperatively to regulate proliferation, specification, and differentiation of midbrain dopaminergic progenitors. J Neurosci. 2011 Aug 31;31(35):12413-25.

Yoshida M, Macklin WB. Oligodendrocyte development and myelination in GFP-transgenic zebrafish. J Neurosci Res. 2005;81:1–8.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):R137.

Appendix 1: Primers used for cloning

Enhancer	Build	Forward primer	Reverse primer
LMX1A -96.34	Hg19	GCAGACGCTATCTCTGCTCTT	CCTGCCATACAAAACCCATT
LMX1A -36.77	Hg19	CCTTCCCTCTCAGTCCTCCT	CAACAATGCCTTTCCACTAGG
LMX1A -8.19	Hg19	TGCAAGAAGAAGTTTGGTGACT	CCAATGGCTGTGCTCCTC
LMX1A -1.59	Hg19	TGCTGCTTTACAGGCATTTTC	TGGTTGTATCTGCCCTATTCC
LMX1A +5.34	Hg19	ACTGATGGGCTAAGCACAGG	GCCCACATAACTTCAAGTATCACC
LMX1A +9.65	Hg19	ATCCATAGTGGCCAGACAGC	TTTGTCTTGATACCCCATAGGA
LMX1A +11.50	Hg19	AGGCGGTAGCCAATAGCAG	CTGTTGCCTCCTCCCTATCA
LMX1A +13.84	Hg19	TTCTGGAAATCAGCCTCCAC	GTCCCCAGCCTCTAACTCCT
LMX1A +17.95	Hg19	TGGATTCCACAGCATGTGAC	TTTCCAATGGCTCTGCTTCT
LMX1A +32.10	Hg19	AAGGTGAGGGGCACAGACTA	AAAGAGCTATGCTCCTTTAGGATT
LMX1A +41.10	Hg19	AACTCCCCTCTCCCAACATC	TGGGATTGTGAAATGCTGTG
LMX1A +67.70	Hg19	CCCAGGAATGCTGCTATTGT	CAAGGAGGCTGGAGTCTCTG
LMX1A +71.89	Hg19	GCGTGAGCTGTAAATGTATGGA	CCCATCTCTCCATGTGCTCT
LMX1A +76.36	Hg19	GGCAAATTTATCCTGCTATTAGTGA	CTTTGGGGTGGGAGACAGT
LMX1A +81.37	Hg19	GTTTTGGGGCATAGAAGCAA	CTCTCCAACAAAGCCCAGAC
LMX1A +92.33	Hg19	TACGCTGAAACAAGCAAGGA	CTGGCTGTAGCATGCTGTGT
LMX1A +102.30	Hg19	CGCCCAGTGTGGTTTTTCT	GGAGGGAAAAACAGAAGGAGA
LMX1A +117.21	Hg19	CTGTGAAGCAACTGCACAAA	TGAGGCTCAGCATTTCTTTTC
LMX1A +123.76	Hg19	GGAGTCCTTTGTCCCTCCTT	ACTTCTGCTGGTCCGTTCAA
LMX1A	Hg19	CAAACTGAGGCCCAAGAA	TGGACGTCTGAGCTGAAG

+135.15		GA	TG
LMX1A	Hg19	TTTGTAGAGTGGCCCAGA	GCAGTTCTTGCCTCCATT
+181.83		GG	GT
LMX1A	Hg19	AGCCCTCTCTTCCTTCCTG	TTTCCCAGGTTGTATTTTT
+184.33		T	ACACA
LMX1A	Hg19	ATTGATCTTTGGGCAGTTG	AGCCCTTTGAGGTCTGGA
+187.31		G	GT
LMX1A	Hg19	TGAGCACCCCATTAGCTCT	TTGGGGAGAGGGAGAGA
+238.85		T	TTT
LMX1A	Hg19	CTGGAAATATTCCCCACTG	CGACCAGTTGTTTGATGC
+296.85		C	TG
LMX1A	Hg19	GGACTTGGGAAGGTCAGT	CAAGGGCCTGTCAACTTT
+475.56		CA	GT
LMX1A	Hg19	CAGAGTGGGCCTATCCAG	CTGGCTGATTCTGGGGTT
+479.56		TC	TA
LMX1A	Hg19	CCATCTCAAATGGTGGCCT	ATATCTCGGGGCGGACAC
+488.50		A	T
LMX1B -93.21	Hg19	GAAGGTCTTGGGGGATGA	CTCGCAGAGGGGGACTC
		TT	A
LMX1B -79.84	Hg19	CTTGGGGTGATGGAGTGA	GCACCTCCCCTCCTGTAC
		GT	TT
LMX1B +3.46	Hg19	CGCCAATTAGGTATCTGCG	ATCAAGAAGGCACGTTTC
		TTA	TG
LMX1B +11.03	Hg19	GTCCTGCTCTGTCCCCTCT	CAGACAGGGCTGGGTCTC
		T	T
LMX1B +12.07	Hg19	GCAGGACGAGTAGAGGGA	GGAACCCTCTGACACCAG
		CA	TC
LMX1B +28.46	Hg19	GTCCAGGGATGATGGATC	GGCTGATACTGACACCAG
		TG	CA
LMX1B +29.70	Hg19	GCTGATAAAGGCCATCTG	CCAGTTATGCGGGAGAA
		GA	GAA
LMX1B +59.40	Hg19	TGCCAGTGTGTGTGTGAGT	CAGGACCACCTTCTGGCT
		G	AC
LMX1B +73.50	Hg19	CCTCCCATACACGCCATTA	TTTGCAGTCAAAGCCACA
		G	GT
Hb_01	Hg19	GGGTGTGAAACTCTTCAGT	CCCACCAACCTGTAATCC
		TTGT	AC
Hb_02	Hg19	TGATGTCATTACGGAACCA	CCTTCTCATTTCTGGCTC
		AG	GT
Hb_03	Hg19	TGGCATCTCACTGATTGTC	GCCTGTTTACACAGAAAA
		TG	CTCTCA
Hb_04	Hg19	CCTTGTGGTGCAATGTGAA	TGCTTAATTCACCCCACT

		T	TTG
Hb_05	Hg19	GAAGAAAGATCTCCCCTTC TCA	CAAATTAACCTATAGCAT CTCAGTGAA
Hb_06	Hg19	ATCAGCAGCTGTAGCGAA CC	GGGAAGGGAAGAACCAC ATT
Hb_07	Hg19	GCATGGATGGTGATCCTTC T	GAGCCAGGACAAGAGTG GAG
Hb_08	Hg19	CCTGTTATCATTGCTATGC TGATT	TGCAAAATTTGGAGGCTT ATG
Hb_09	Hg19	CATGATAAAAACAGAAGA AGCAAAA	TGAAATGAGGACACATG CAAA
Hb_10	Hg19	GTGGAAAGTCTATCTCCAT GTCG	TGGCACTAAGCTGCTTTT GTT
Hb_11	Hg19	TGAATATCTCAACTGAATT TTCTTAAA	TTGCTTTTCTTGATGTTTT TCG
Hb_12	Hg19	CAAACTGGGGAAACTTGC TT	CCGGTACCGACTCTTTTC G
Hb_13	Hg19	CGATGGCAGGGGACTAAG TA	CCGATAAGCCCCTACTCT GA
Hb_14	Hg19	GGCACTGAGGGTGTAGAA GC	CTACCCCAGCAGCAAATG AT
Hb_15	Hg19	CTGAATCCAGAGCCATCCT C	ATTCCTCCCAGCTCTCTG CT
Hb_16	Hg19	CATATATGAATGTATGAG GTTTTCCA	AAAAGAGGTGTCCTTGTT CTGC
Hb_17	Hg19	TAGGCTCACTGCCTGCTCT T	AGGAGACCAGCCACGTG TTA
Hb_18	Hg19	TGCCTATTGGCATAGTTAG GTTG	CACTTTCTTTGGTTTTCAT GTGTT
Hb_19	Hg19	GCATACTTCATGGTGCCTG T	TTGTCAGTGGCTGCACAA GT
Hb_20	Hg19	TCCTCCAACACAAAATTCA AAA	TAACGTGCTGGCATGCTG
Hb_21	Hg19	ACGAGGAAGGAGATTGAG CA	TCCAGAGGAAAGCTGGA AAG
Hb_22	Hg19	CCACCCCTCAAAGTGCTAT T	AAATAAAGGCACCTAGT ACCATGC
Hb_23	Hg19	TCACATTGACTGTGTTTAC ATCCA	GCAGTCTCAGGGGAGGA CTT
Hb_24	Hg19	AGAGCTGGAGCTGGGAGG TA	GGGAGAGGATGGGGACA TAA
Hb_25	Hg19	CAGGATGCTTCAGAAATG	AAAATGTTACGGGAAGC

		ACAG	AACA
Hb_26	Hg19	CTTTCCTTCTGGGCTGAGTG	TGGTTGCCTGGCGTTTAGTA
Hb_27	Hg19	TGCCAGTAAACTCTGGTTTTCA	TGTGTTCCCCTGAATCTGCT
Hb_28	Hg19	GCTTGCAAACCTGGCTTATT	GGACACTGAGGAACCCTTGA
Hb_29	Hg19	TCTAGCTTTATTAGAAAATGTGAAACA	GGCAGGAAGATTACGAATGG
Hb_30	Hg19	TTTCCAGCCTCATTCTAGGC	TTCCCTTTGAAGCCAAGCTA
Hb_31	Hg19	TTTGATGACTTAAGAAAACTGCAT	TATCCGCCGAGTAAACCAAC
Hb_32	Hg19	TTAGGGGGATTGATGGTTTC	AAATTATCTTCGGGGAAGTTTT
Hb_33	Hg19	TGTGTTCTTTCCCTGGGTTT	ACTCCCACAGACACTGTTCC
Hb_34	Hg19	TCCATGACTTCAAAACTATTTTATAGCA	AACAGCAGACTCTATTTTCATGC
Hb_35	Hg19	CCTATTTATGCAAATGGTTGGA	CCTGCTGTGGACTGTCTTGA
Hb_36	Hg19	AAAACGGGAGAGGTCATGGT	GTTTTCCCGGACCAACATTT
Hb_37	Hg19	AAAACTAGAATGCTCATTTCTTTCA	AGGAGGCATCTTAGAAGTCCA
Hb_38	Hg19	GGCGAATTAAATAAATATTCCTGTTG	CAGTTGTGGTGCAAAAGCAG
Hb_39	Hg19	AGCTCGTTGGCCTCTCTATG	GGTTGGAGGCTGCTCTCTC
Hb_40	Hg19	GACCTCCCGTGGGACTTG	CCGGTAACTTTTTCCACACG
Hb_41	Hg19	CCTTCTAGCCCCCATCATTT	GGATGCCTGCTTAACTCCAG
Hb_42	Hg19	TGCAAAAAGAAAAACAGGGATT	ACAAACCATTATGTAACTTTTACGAA
Hb_43	Hg19	AATTGGTAAAGGCTGTGAGCA	GGATCAAGTTTATTCCACATCTTCA
Hb_44	Hg19	GCAGAGTCTACCCTCCTGGAC	GTGACAGGTGGATGGTGGA
Hb_45	Hg19	GCATTTGCACAGAGGATCG	TGTGGCTTTCTAGGCTGGTT
Hb_46	Hg19	CAAGGTGCAACGAGATAC	CTCCTTAACCCCCATCAC

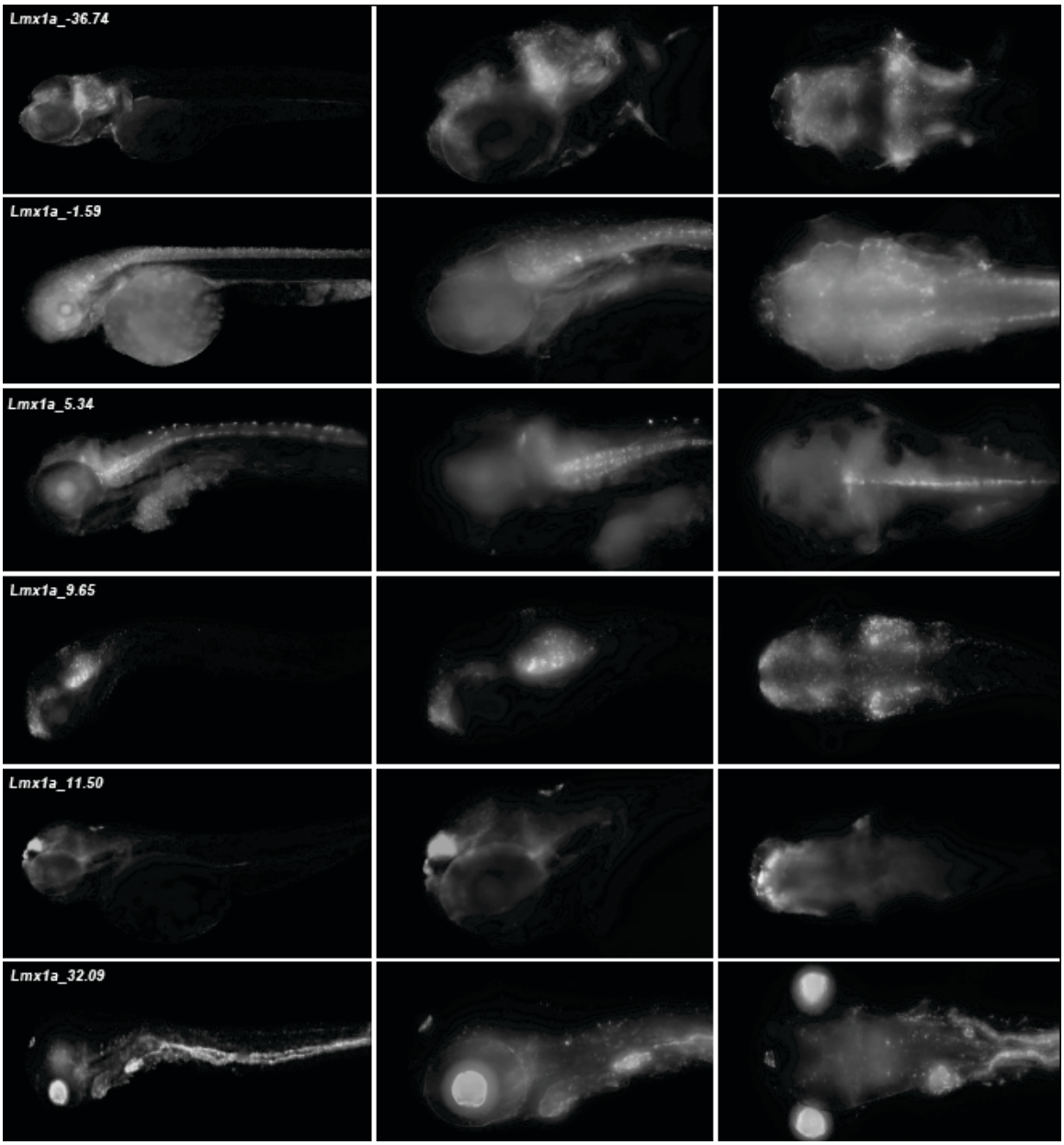
		CA	CT
Hb_47	Hg19	TGAAACTTCGTACCTCATT AACAAA	AAAACTTCTTTCTGCCGT GAA
Hb_48	Hg19	CCCAGTAAGCGTATGGTA ACTTT	CAATGGATCACTGCTTTA CAGTTT
Hb_49	Hg19	CATTAACGTTGACTGAGTT CTTGG	AGGGACTTCTACTTGGAC TCTCA
Hb_50	Hg19	CAAATATGGGAGATCCAG GAA	TTTTAGAATTCTGTCCCC AACA
Hb_51	Hg19	CCTTCAGCATAAAACAGT GACC	CGATTGAAAACATGTAAT TTGAGC
Hb_52	Hg19	CCAGGGGAGTGACACATT TT	AGAAGAAGAGCAGGCCC AAC
Hb_53	Hg19	CGTTTCTCCCAAGAGCAGA C	GTCTCCCCTCTTTCGGTTT T
Hb_54	Hg19	CGTACCACCACACTTGGCT A	GGACTTATTACAAAACAC GCAAAA
Hb_55	Hg19	CTTGCACCACATATCCAGG TT	TACCATGGGCTGAGGAA AAG
Hb_n01	Hg19	TCCATTAGAGGCACCTTGC T	GGGCTACTTTTGAAGCAC CA
Hb_n02	Hg19	TGATGGGGTTTCTACCTTA CCA	AACCCATAGAATGCGCA GAG
Hb_n03	Hg19	AGAGACGGGGTTTCATTGT G	ATCCCCCTTTGTCCTTGA AC
Hb_n04	Hg19	CTTCTCCCACCGGCTCTC	GGTTGGTCGTCTGGGTAC AG
Hb_n05	Hg19	TTCACACTCTGTGGTGACT GC	CATACAATGGGACCCTCA GC
Hb_n06	Hg19	TGAGAAACATTCAGTGCA GAAA	AACTGAGTCCCCAGGAG GAG
MelanA-UF1	Mm9	ACCTCTCCACAAGCCACAT T	CCAACTTCCCAAGGTCAG TG
MelanA-UF2	Mm9	TTCCTAGACTTCCGGGGAT T	CACCATGTTCTTTGTCGG AAA
MelanA-UF3	Mm9	GACTGAGCGAGTGTGATG GA	TTGACATCCACCCCTTTT TC
MelanA-UF4	Mm9	CTCACATCCTGCACATCTT TG	TGGTAAGAGGCTCAGCA AGG
MelanA-UF5	Mm9	TTTGAATACACAAACAGA CTTGGTT	CCAAGTGTCTGAAGTGAT AGGTTT
MelanA-UF6	Mm9	TGACCCTCACTTGGCACAT	CTTCCTGGGGTCAGTTCA

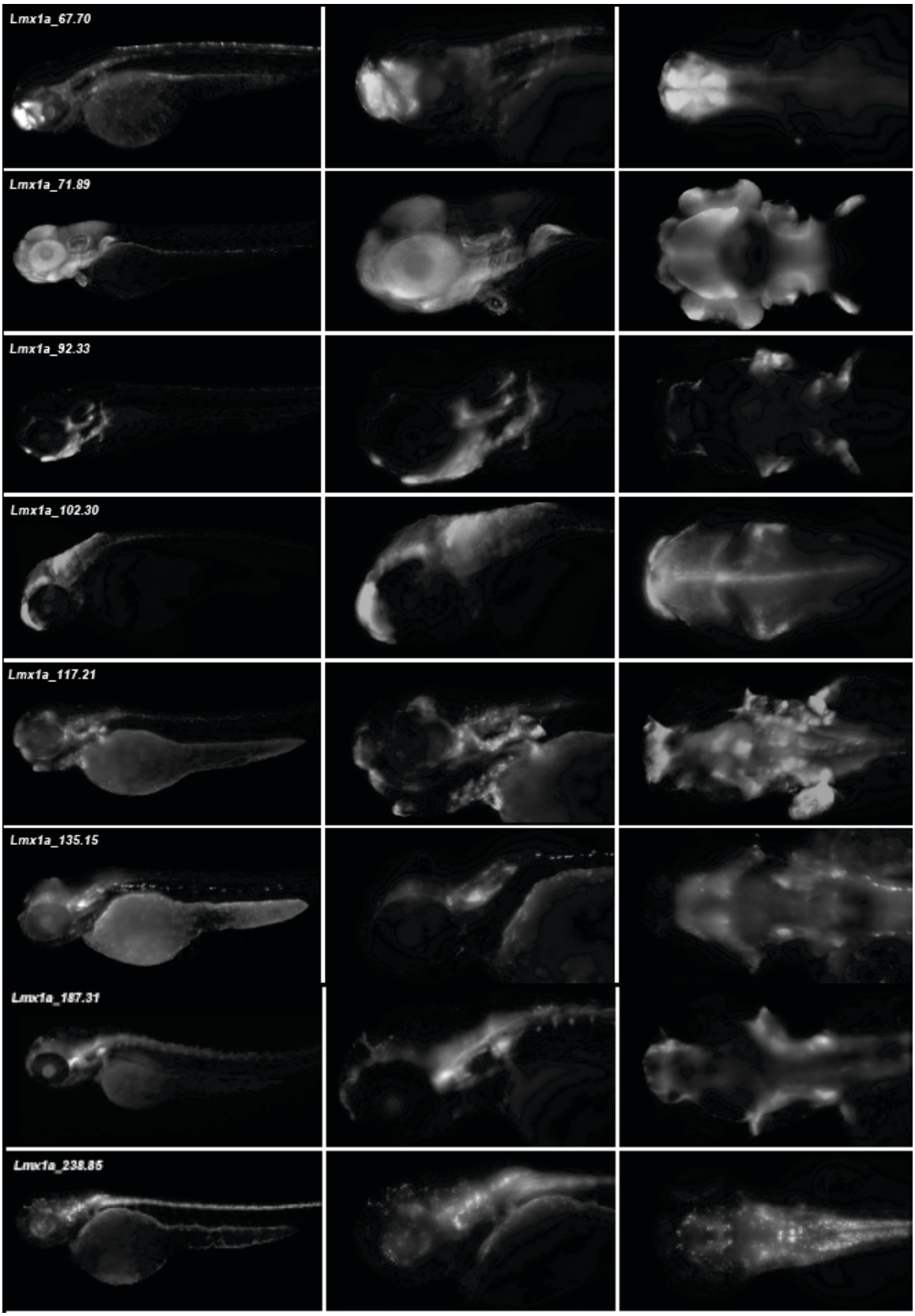
			AA
MelanA-UF7	Mm9	CCCAAAAGAAGAAAAAGCACA	TTCAGAAGGAATTTACCCCATC
MelanA-Cons1	Mm9	TGGTGATTCTCTTTGACCACTG	GGAGCAGGATCCCACAAATA
MelanA-Cons2	Mm9	CCAGTCTGGCATTCTTCAGG	TGCGAGTCTAGATAGCTCAATTATACA
MelanA-Cons3	Mm9	TCGCAACACTATTAGCAAGAGC	AGGGTTTAACGTGGGGTTTC
MelanA-Cons4	Mm9	AGAGAAAGCCAGAGCACCTG	AACCTGCCACCACTGTTAGAA
MelanA-Cons5	Mm9	CAGTGGAAATGGCAAAGAGTT	GCTCATTTGTTACCATGACCAG
MelanA-Cons6	Mm9	GCATCATGCCACCACATTAC	CCCAAAAATTAGAAGAGGAAGC
MelanA-Cons7	Mm9	TGATGATTGGCCATTTGTGT	CCTAGTAAGGGCCACAGCAC
MelanA-Cons8	Mm9	AGCCCAGGGAATTTATGCTT	GAAGGGAGGGGGAAATAACA
LMX1A +1.43	Hg19	ATCTCTGGCTTCAGCTCTGC	GGGTCGTCCCTACCCAAATA
PITX3 -0.004	Hg19	CCCATTCACTTTATGGCAGAC	GTGGGAGGATTGGAACAGAG
PAX2 -1.02	Hg19	GACGGGGGCTGGAGGAATCT	CTGAACCCTCGGGTGACTGG
SLC18A2 -23.88	Hg19	CTGGAGCTGATGGAGGAATC	CTTGGACCTGGCTACTGAGC
LMX1B -2.90	Hg19	GACCAGCAGAGGGCAGAAAG	GTGTGTGTCCGCATCATTTC
EN2 -76.15	Hg19	GACTTTTGCAGCCTGGCTCT	CCAGAGTTGGTGGCTCTGAC
GBA +1.38	Hg19	TGGCTGTTTGACTTTCAATAAATC	GCTCAGTACCTGGCCAAAAA
ARHGEF2 +10.54	Hg19	AACTCAGGCAAATCCACAGC	TGTGAGGATGAGGAGGGAAC
SUFU +92.22	Hg19	CCAGTGTGGTACCTGTGGAG	AGAACCTCCCAGTCACTGCT
PROX1 +18.01	Hg19	CAGCAAACTGGGCAAGATG	CTCCAGGTGGCTGTGTTTC
PBX1 +77.40	Hg19	GCCACTTAGAAGGCCTGAGTT	CAAGGGCATATGCATTTTCTT
PBX1 +78.50	Hg19	CAGTGGATTCTTCAGAGGT	CTATGGCTAGGGCAGAG

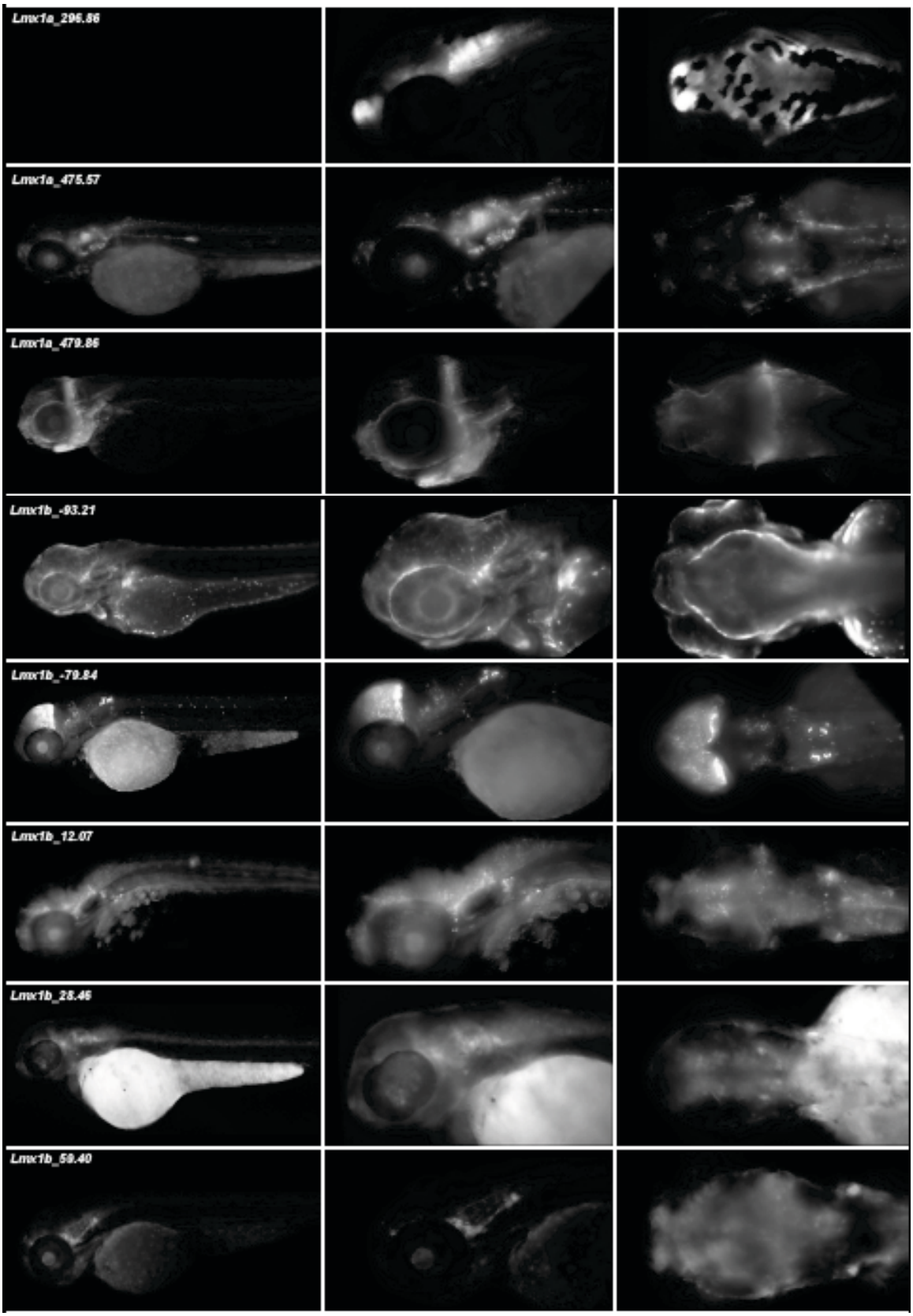
		ATGG	AGC
POU3F3 -18.90	Hg19	TCACAACTGTTTTGGATGT CTTG	GCCATGTAATTTCTTTAT CTTTCACC
SOX2 +7.01	Hg19	CCTGCTAACAGTGCTGTCC A	CTGGGAGAGCTAGGCAA ATG
SOX2 +18.01	Hg19	TTTGTATGAACTGAGCACA GAAGG	CCTTCTACATTCCACCAC TGC
POU3F2 -191.20	Hg19	CTCATGTTTTCACTCAAGT GCT	TGGTTGCTAACCAGAGGA GAA
POU3F2 -191.90	Hg19	CATAATTTGAATTGTGATT TGCAC	CCCAAAGGTAATTATTGC TAATGC
ETV1 -204.80	Hg19	GGGACTGTTATTGTCGAAA GC	TTGCAAGTTCTCTCTTCT AAGTGATT
NKX2-1 +1.32	Hg19	GATTTCAGCCCTCTCCTTG A	GAAAGAAGCTGGGTGAC TGG
Sox2 -7.48	Mm9	CCATGCTCGGTTGCATTGT G	AAAGCTCATCCTCGAACC TGG
Atp13a2 +24.96	Mm9	TGCGCTCATCCTGGACATA G	CAATCCTGTCTACCCCTG GC
Dnajc6 +.0631	Mm9	TCCGAAGGGAACAAGATA GGC	AACGAAGAGGTTACAGG TGGC
En1 +213.76	Mm9	TCCTCTTTATGCTGTCTCA AGAA	AAATCCAGGCGACAACT GCT
Shh +2.77	Mm9	CAGAAAGGCTACTGCCCA GG	TAGACGCCTTTCAGCTCC AC
Nr4a2 prom	Mm9	CCGCGCTCGCTTTGGT	GCCTGACCTCTCATCCTT CG
Rxrg -99.82	Mm9	GGAGTCAGGAAATACAGA AGGAAAC	TCTGAATACAGAAGGGCT GGT
Pink1 +0.783	Mm9	GTCCCCCTTTTATGCATCC CA	GAACCGGGCCAGAGTGT AG
Gbf1 +25.07	Mm9	ACAGGCAGTGATTCCCCTT TT	TTCCGGGATATGGAACAT GTGG
Th -7.70	Mm9	TGGTGTCTTAGGGAAGGG CT	GAGTCCAGCCACCTTCTG TC
Th -18.01	Mm9	GGGTCTGAGTTTCAGGAG GC	GAAAGAGGATGGGCAGC AGT
Th -18.55	Mm9	ACTGCTGCCCATCCTCTTT C	ACCCAGACACTTGTCCCT CT
Otx2 -15.02	Mm9	TGTTTGCTAGCCATGTTTT CCTT	CTGGAAAATCCCCGCCAG AC
Pax2 +51.28	Mm9	TGTTTGCTAGCCATGTTTT	CTGGAAAATCCCCGCCAG

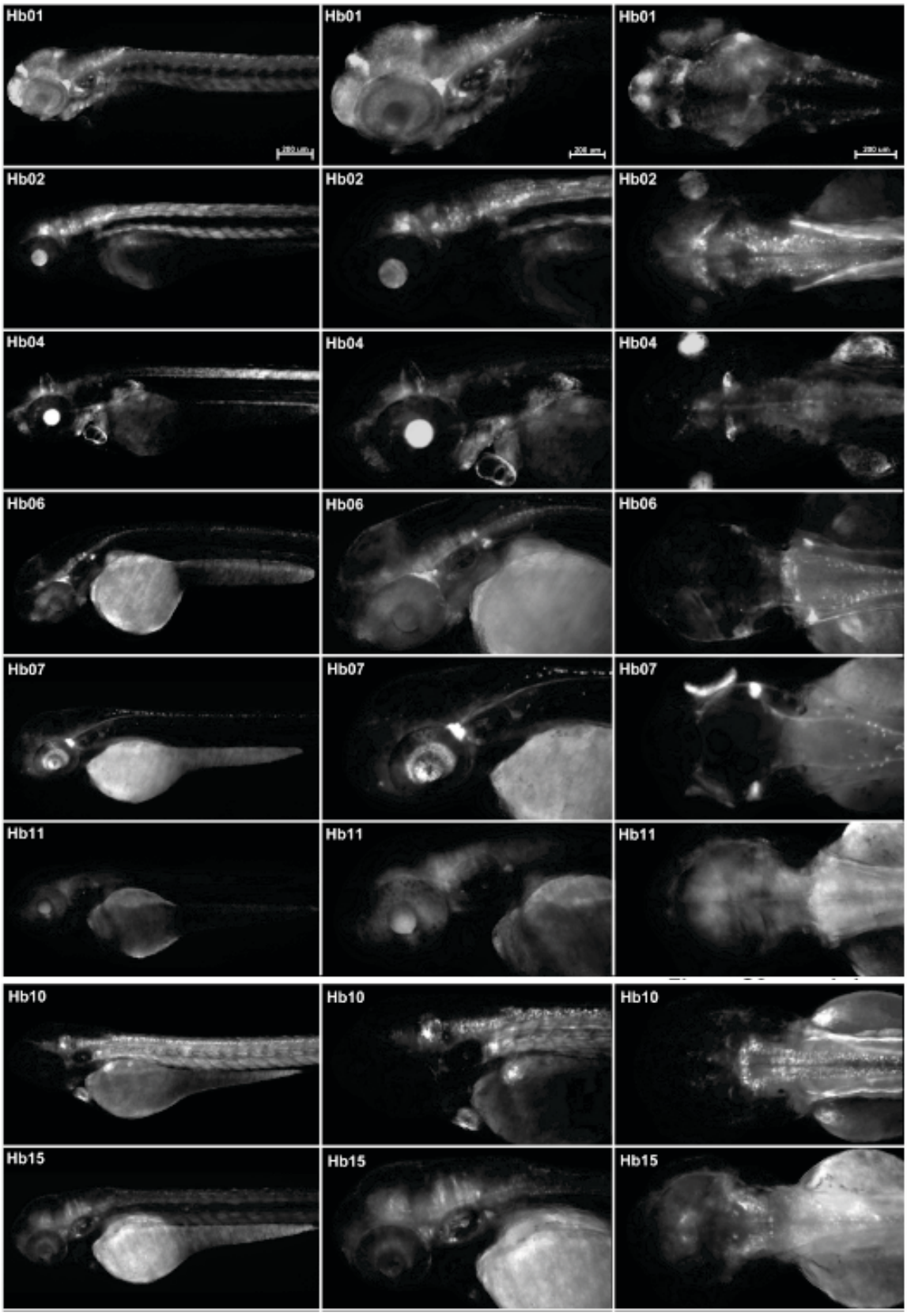
		CCTT	AC
Park2 +226.94	Mm9	GTAGCCAGGGCCAAAGAC CC	CTACGCCCCCGAGATACA AAG
Park2 +514.83	Mm9	ATTTCCCAGCTTCCGGTGT T	GTGTTTGCTGCCCAAGAG TG
Vax1 +40.40	Mm9	CATACCCCCTCATCCTATG CTT	TGAGTCCCTGCCCTTTGA AATC

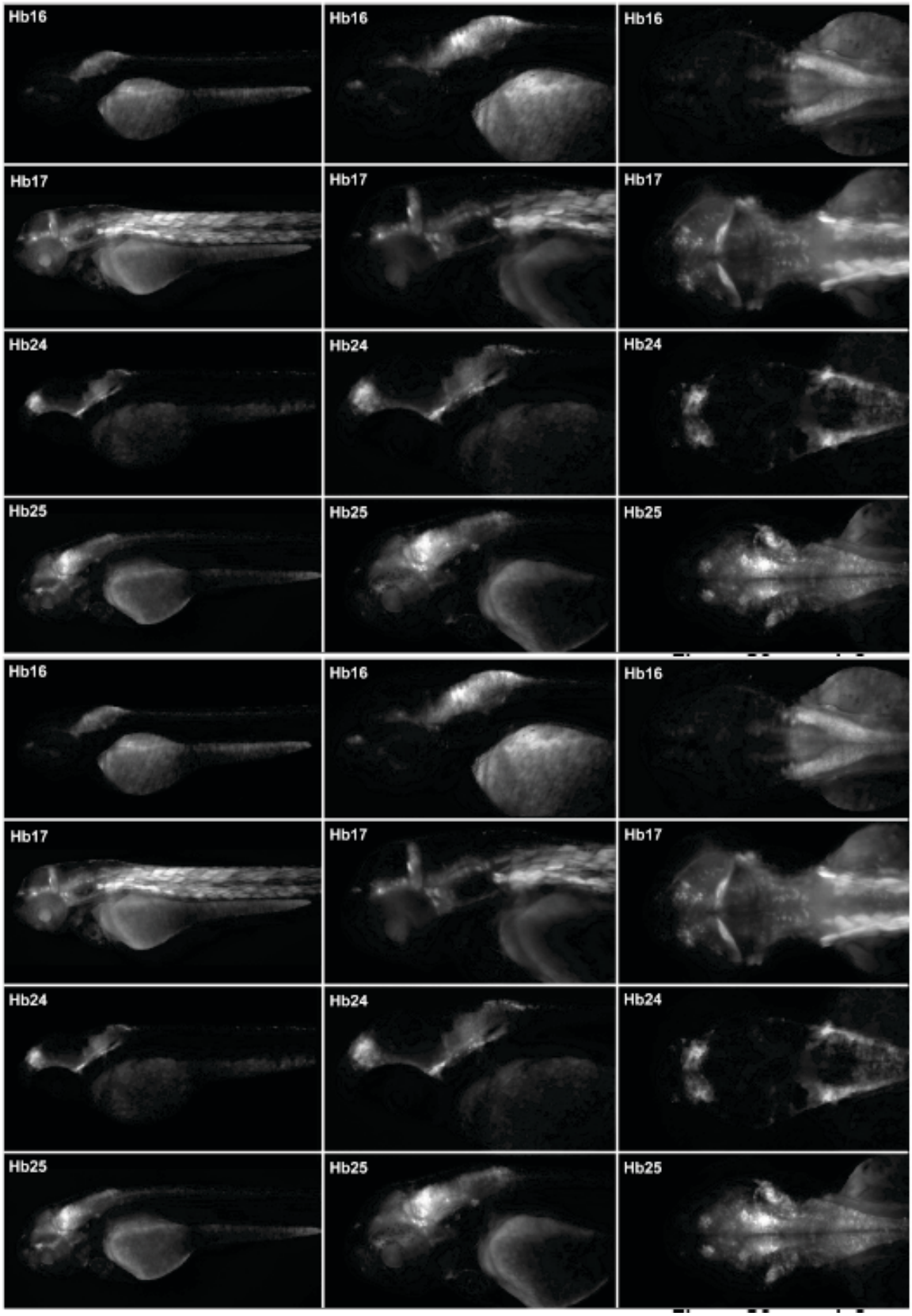
Appendix 2: Representative EGFP expression patterns driven by all elements with multiple zebrafish founders.

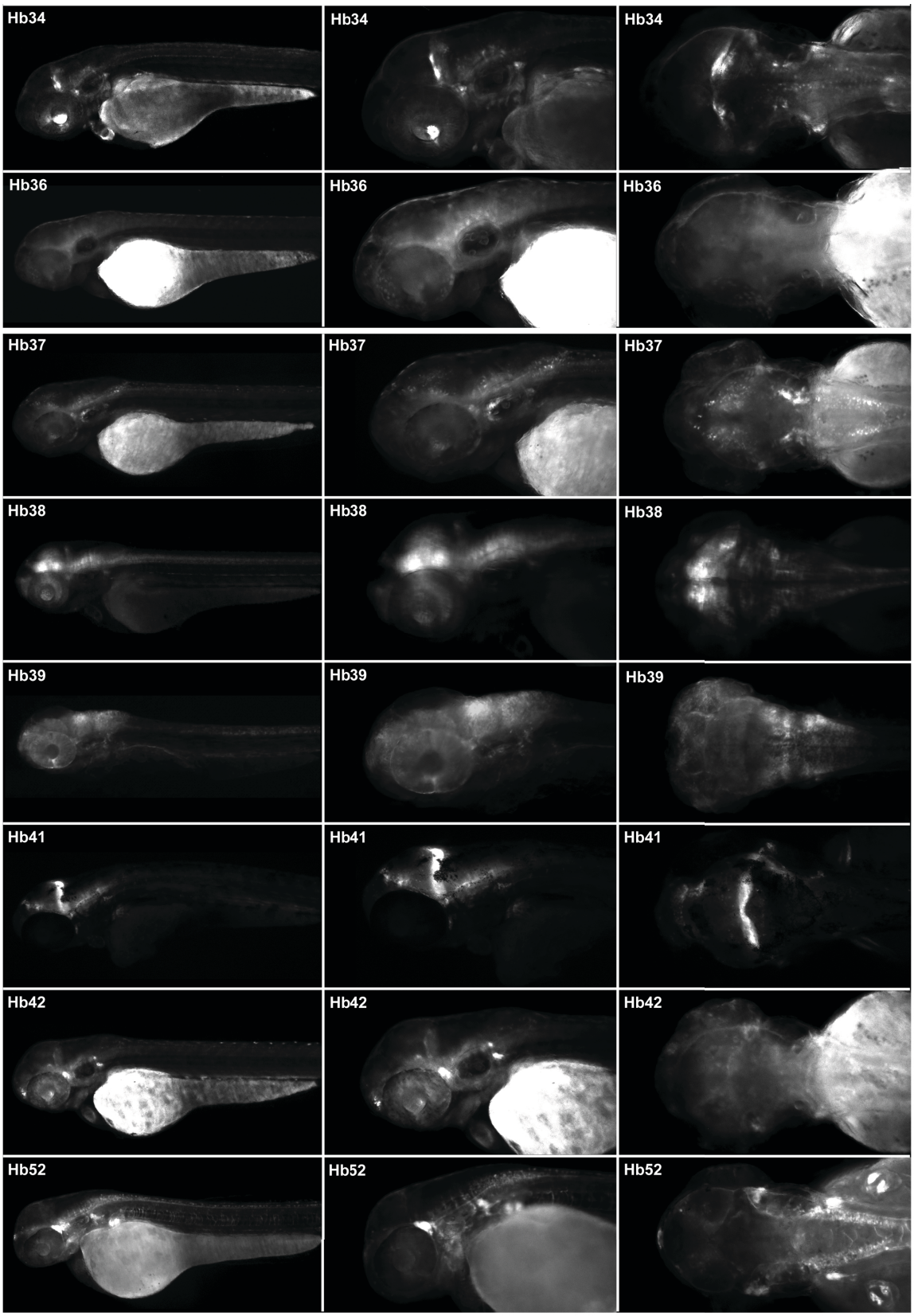


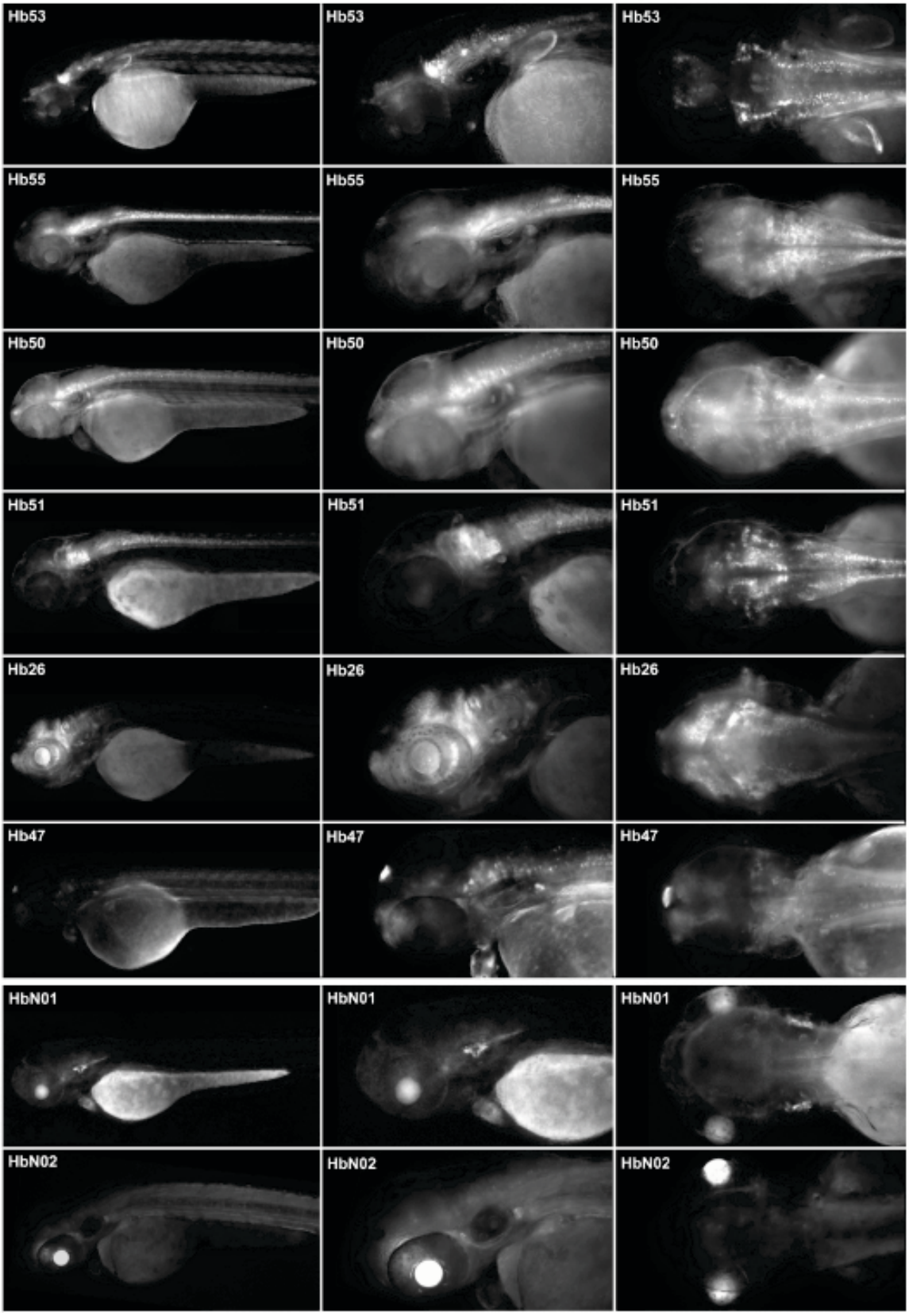

Lmx1a_67.70
Lmx1a_71.89
Lmx1a_92.33
Lmx1a_102.30
Lmx1a_117.21
Lmx1a_135.15
Lmx1a_187.31
Lmx1a_238.85

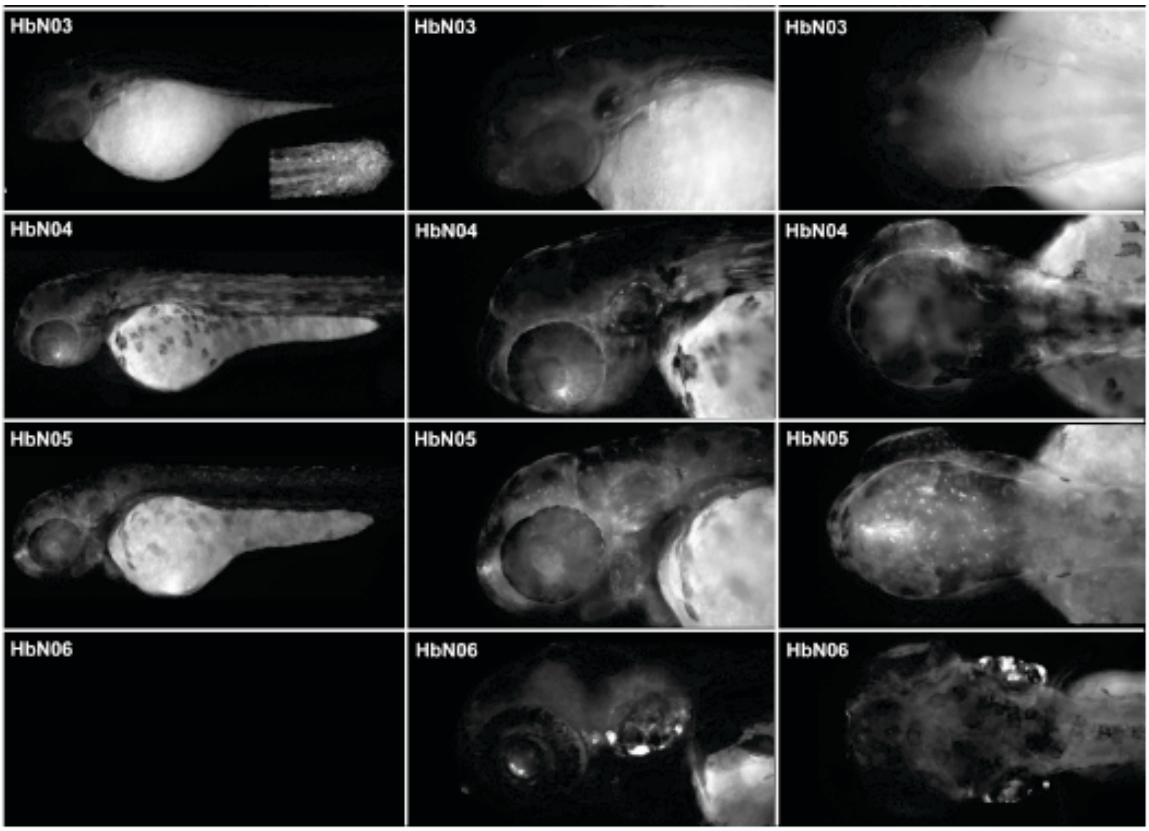

Lmx1a_296.86
Lmx1a_475.57
Lmx1a_479.86
Lmx1b_-93.21
Lmx1b_-79.84
Lmx1b_12.07
Lmx1b_28.46
Lmx1b_59.40


Hb01
Hb01
Hb01
Hb02
Hb02
Hb02
Hb04
Hb04
Hb04
Hb06
Hb06
Hb06
Hb07
Hb07
Hb07
Hb11
Hb11
Hb11
Hb10
Hb10
Hb10
Hb15
Hb15
Hb15


Hb16
Hb16
Hb16
Hb17
Hb17
Hb17
Hb24
Hb24
Hb24
Hb25
Hb25
Hb25
Hb16
Hb16
Hb16
Hb17
Hb17
Hb17
Hb24
Hb24
Hb24
Hb25
Hb25
Hb25


Hb34
Hb34
Hb34
Hb36
Hb36
Hb36
Hb37
Hb37
Hb37
Hb38
Hb38
Hb38
Hb39
Hb39
Hb39
Hb41
Hb41
Hb41
Hb42
Hb42
Hb42
Hb52
Hb52
Hb52


Hb53
Hb53
Hb53
Hb55
Hb55
Hb55
Hb50
Hb50
Hb50
Hb51
Hb51
Hb51
Hb26
Hb26
Hb26
Hb47
Hb47
Hb47
HbN01
HbN01
HbN01
HbN02
HbN02
HbN02


HbN03
HbN03
HbN03
HbN04
HbN04
HbN04
HbN05
HbN05
HbN05
HbN06
HbN06
HbN06

Appendix 3

Mol Genet Genomics (2013) 288:579–589
DOI 10.1007/s00438-013-0771-7

ORIGINAL PAPER

Integration of genomic and functional approaches reveals enhancers at *LMX1A* and *LMX1B*

Grzegorz M. Burzynski · Xylena Reed · Samantha Maragh · Takeshi Matsui · Andrew S. McCallion

Received: 21 August 2012 / Accepted: 25 July 2013 / Published online: 13 August 2013

Abstract *LMX1A* and *LMX1B* encode two closely related members of the LIM homeobox family of transcription factors. These genes play significant, and frequently overlapping, roles in the development of many structures in the nervous system, including the cerebellum, hindbrain, spinal cord roof plate, sensory systems and dopaminergic midbrain neurons. Little is known about the *cis*-acting regulatory elements (REs) that dictate their temporal and spatial expression or about the regulatory landscape surrounding them. The availability of comparative sequence data and the advent of genomic technologies such as ChIP-seq have revolutionized our capacity to identify regulatory sequences like enhancers. Despite this wealth of data, the vast majority of loci lack any significant in vivo functional exploration of their non-coding regions. We have completed a significant functional screen of conserved non-coding sequences (putative REs) scattered across these critical human loci, assaying the temporal and spatial control using zebrafish transgenesis. We first identify and describe the *LMX1A* paralogs *lmx1a* and *lmx1a-like*, comparing their expression during embryogenesis with that in mammals, along with *lmx1ba* and *lmx1bb genes*. Consistent with their prominent neuronal expression, 47/71 sequences selected within and flanking *LMX1A* and *LMX1B* exert spatial control of reporter expression in the central nervous system (CNS) of mosaic zebrafish embryos. Upon germline transmission, we identify CNS reporter expression in multiple independent founders for 22 constructs (*LMX1A*, $n = 17$; *LMX1B*, $n = 5$). The identified enhancers display significant overlap in their spatial control and represent only a fraction of the conserved non-coding sequences at these critical genes. Our data reveal the abundance of regulatory instruction located near these developmentally important genes.

Keywords Enhancer · LMX1A · LMX1B · Midbrain · Hindbrain · Zebrafish

G. M. Burzynski and X. Reed contributed equally to this work.

Communicated by T. S. Becker.

Electronic supplementary material The online version of this article (doi:10.1007/s00438-013-0771-7) contains supplementary material, which is available to authorized users.

G. M. Burzynski · X. Reed · S. Maragh · T. Matsui · A. S. McCallion (✉)
McKusick-Nathans Institute of Genetic Medicine, Department of Molecular and Comparative Pathobiology, The Johns Hopkins University School of Medicine, 733 N. Broadway, BRB Suite 407, Baltimore, MD 21205, USA
e-mail: andy@jhmi.edu

X. Reed · S. Maragh
Predoctoral Training Program in Human Genetics, McKusick-Nathans Institute of Genetic Medicine, The Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA

S. Maragh
Biochemical Science Division, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA

Introduction

LMX1A and *LMX1B* encode closely related transcription factors (TFs) that contain LIM-homeodomain (LIM-HD) motifs. They play pivotal roles during nervous system development, specifically in neural tube regionalization, the extension of axonal projections and the acquisition of neurotransmitter phenotypes (Dai et al. 2009; Hobert and Westphal 2000; Shirasaki and Pfaff 2002). Despite their clinical and developmental importance and the significant

inquiry that these *LMX1* genes have been subject to, relatively little is known about the sequences that establish the genomic regulatory landscapes required to execute their developmental programs. We set out to provide a significant, though incomplete, characterization of the *cis*-regulatory landscape of the *LMX1A* and *LMX1B* gene intervals by identifying elements that display CNS regulatory control. Such CNS enhancers are considered candidates for *LMX1A* and *LMX1B* transcriptional control and thus also may contain variation therein that could compromise their function and underlie disease risk.

Both *LMX1A* and *LMX1B* are involved in hindbrain and spinal cord roof plate formation and in directing the development of midbrain dopaminergic (DA) neurons (Mishima et al. 2009; Nakatani et al. 2010; Yan et al. 2011). *LMX1A* is an essential regulator of neuronal proliferation and differentiation in the cerebellar rhombic lip and telencephalic cortical hem (Mishima et al. 2009; Chizhikov et al. 2010). In addition, *LMX1B* plays a role in formation and function of the isthmic organizer (IsO), which directs the establishment of midbrain and hindbrain regional identities (Adams et al. 2000; Guo et al. 2007). It has also been shown to be necessary for serotonergic neuronal specification (Ding et al. 2003). Studies in mice first established the impact of *Lmx1b* deficiency, and lead to the demonstration that *LMX1B* mutations were responsible for human nail patella syndrome (Chen et al. 2003; Vollrath et al. 1998). Similarly, *Lmx1a* mutations were initially described in mouse neurological mutant *dreher*, which displays defects in cerebellar, hippocampal, and cortical development, as well as hindbrain roof plate malformations, short tail and deafness consistent with the patterns of its embryonic expression (Millonig et al. 2000; Failli et al. 2002).

Instructions encrypted within transcriptional regulatory elements (REs) such as enhancers instruct cell fate determination, and render cells transcriptionally competent to respond to their environment (Noonan and McCallion 2011). Although regulatory variation is expected to contribute significantly to disease risk (Kleinjan and Coutinho 2009; Visel et al. 2009; Maurano et al. 2012), REs, unlike coding sequences, lack an established vocabulary to facilitate their immediate recognition in primary sequence. The recent emergence of chromatin immune precipitation (ChIP)-based strategies coupled to next generation sequencing (ChIP-seq) has facilitated the identification of REs in a sequence agnostic manner. However, these approaches may not be well suited to comprehensive analyses of single genes, particularly those with pleiotropic expression in discrete cell populations that cannot be obtained in sufficient numbers. In situations such as this, evolutionary sequence conservation still provides a powerful tool for the identification of functional sequences, and although conservation alone is unable to discern the biological roles of sequences, one can, through functional analyses, reveal REs with a wide range of regulatory control (Noonan and McCallion 2011). When available, sequence intervals may be cross-referenced with pertinent ChIP-seq/DNase-seq data to provide additional evidence of regulatory activity and help predict cell-type-dependent activity.

We selected 71 human, conserved non-coding DNA regions at *LMX1A* and *LMX1B* for preliminary functional evaluation using transgenesis in zebrafish. Of these sequences, 47 (66 %) directed reporter expression in the central nervous system (CNS) of mosaic G0 embryos. We identified multiple independent founders for 22/45 enhancers at *LMX1A* ($n = 17$) and *LMX1B* ($n = 5$). Each directs expression in aspects of the developing nervous system of zebrafish embryos, consistent with expression of their respective endogenous genes. All 22 enhancers directed reporter expression in the CNS. A subset of these enhancers direct expression in the diencephalon, the cerebellum and at the midbrain–hindbrain (Mb–Hb) boundary, consistent with the critical role of LMX1 factors in the development of hindbrain roof plate and IsO formation. Many also direct expression in the peripheral nervous system (14/22) and non-neuronal tissues such as the otic vesicles, cartilage, pronephros, and muscles. This study adds significantly to the number of enhancer elements identified at *LMX1A* and *LMX1B*, but perhaps more importantly, it reveals that the complexity of regulatory control can exist at individual loci.

Results

Evolutionary conservation facilitates identification of zebrafish lmx1a and lmx1b genes

Evaluation of putative *LMX1A* and *LMX1B* regulatory sequences in a zebrafish model is aided by an appreciation of the spatial expression of their teleost paralogs. Thus, we first set out to identify zebrafish *LMX1A* and *LMX1B* paralogs. Approximately 30 % of the gene content of *Danio rerio* remains duplicated subsequent to an ancient genome duplication event in the teleost fish lineage (Amores et al. 1998). The zebrafish genome contains two identified *LMX1B* paralogs (*lmx1ba* and *lmx1bb*). However, only one *LMX1A* paralog (*lmx1a*) had been identified in the zebrafish genome at the time of these experiments (http://www.ensembl.org/Danio_rerio/Gene/Summary?g=ENSDARG00000020354;r=20:33946947-33964868, Zv9). We performed a BLASTP query of the zebrafish peptide database in GenBank using the human *LMX1A* RNA sequence (NM_001174069.1) and identified another potential paralog previously annotated with 'predicted' status (LIM homeobox transcription factor 1-alpha-like,

XP_001922131.3). LMX1A displays 66 % identity to LMX1B at the amino acid level, and is 58 and 59 % identical to zebrafish Lmx1a and Lmx1a-like, respectively. LMX1B paralogs are even more similar; Lmx1ba is 72 % identical and Lmx1bb is 82 % identical to the human LMX1B protein sequence (NP_001167617.1). Figure S1 provides a phylogram illustrating the similarity among the amino acid sequences that encode LMX1A, LMX1B and their zebrafish paralogs. The paralogs of *LMX1A* cluster together, but in a distinct node from their human counterpart. By contrast, *LMX1B* shares a common node with its zebrafish paralogs.

Zebrafish lmx1 genes are expressed throughout the central nervous system

We performed whole mount in situ hybridizations (ISH) to document the spatial and temporal expression patterns of the endogenous *lmx1a*, *lmx1a-like*, *lmx1ba*, and *lmx1bb* genes, and to determine the level of similarity to the published expression of their mammalian orthologs in mice. Aspects of the early developmental expression of *lmx1ba* and *lmx1bb* (formerly called *lmx1b.2* and *lmxb.1*, respectively) have been previously described (O'Hara et al. 2005; Cheng et al. 2007; Elsen et al. 2008; McMahon et al. 2009; Filippi et al. 2010). We present their analysis here to facilitate comparison with expression of *lmx1a* and *lmx1a-like*.

lmx1a expression in the CNS was diffuse, broadly overlapping and extending beyond *lmx1ba* and *lmx1bb* expression domains (Fig. 1). We detected transcript from *lmx1a* throughout the brain, including the diencephalon and telencephalon, at both 48 hours post fertilization (hpf) and 72 hpf. In addition, we detected more localized signal corresponding to *lmx1a* in the ventral diencephalon, raphe nuclei and otic vesicles at both time points, and at 72 hpf saw specific labeling of the cranial ganglia (Fig. 1a–d). By contrast, *lmx1a-like* expression was regionally restricted, with distinct labeling of the epiphysis, ventral diencephalon, rhombic lip, and raphe nuclei, closely resembling expression of *lmx1ba* (Fig. 1e–l).

We detected expression of *lmx1a-like* in the anterio-dorsolateral hindbrain and in the ventro-midline, corresponding with the cerebellar rhombic lip and serotonergic raphe nuclei, respectively (Fig. 1e–h). Both *lmx1a* and *lmx1a-like* appear to be more highly expressed in the anterior raphe nuclei at 48 hpf (Fig. 1a, b, e, f), while both *lmx1b* transcripts are detected approximately equally in both raphe nuclei populations (Fig. 1i, j, l, m). These data are consistent with both *Lmx1a* and *Lmx1b* mammalian counterparts, which are also expressed in the developing cerebellum and serotonergic neurons. *Lmx1a-like* has very little dorsal hindbrain expression at 48 hpf, but by 72 hpf, transcripts are detected strongly in the posterior dorsal hindbrain (Fig. 1g, h). This pattern is unique to *lmx1a-like* while some expression domains overlap expression of *lmx1ba* and *lmx1bb* in the ventral diencephalon, rhombic lip, serotonergic raphe nuclei and faintly in the otic vesicles (Fig. 1i, j, m, n) (Cheng et al. 2007; Filippi et al. 2010). Notably, strong expression is seen for all transcripts in the ventral diencephalon through 72 hpf, the area where main clusters of DA neurons are formed, consistent with their role in the induction of midbrain DA neurons (Mishima et al. 2009; Yan et al. 2011).

The patterns of expression are similar between *lmx1ba* and *lmx1bb* with common domains in the ventral diencephalon, raphe nuclei, rhombic lip, and dorsal hindbrain, as well as the amacrine neurons of the retina at 72 hpf (Fig. 1i–p). *Lmx1bb*, and to a lesser extent *lmx1ba*, show additional expression in the dorsal diencephalon that is not seen for the *lmx1a* transcripts (Fig. 1i–p). Overall, *lmx1bb* shows broader domains of expression than *lmx1ba* throughout the CNS; however, *lmx1ba* transcript is also unexpectedly detected in the heart (Fig. 1i–l). All probes were designed to exclude the potential for cross-hybridization with other lmx1 family members, and with other unrelated transcripts (Methods).

Selection of non-coding sequences at human LMX1A and LMX1B genomic loci

The human *LMX1A* gene comprises seven exons, encompassing 154 kb on chromosome 1q24 and is flanked by *PBX1* and *RXRG* (Fig. 2a). *LMX1B* includes eight exons that encompass 112 kb of chromosome 9q33.3 and is flanked by *FAM125B* and *ZBTB43* (Fig. 2b). Candidate intervals for functional analysis were selected from the sequence contained between their respective flanking genes, therefore providing *LMX1A* and *LMX1B* regions of 554 kb and 298 kb, respectively. Although this search space was not exhaustively explored, we prioritized conserved non-coding sequences for assay using genomic evolutionary rate profiling (GERP) (Cooper et al. 2005), successfully PCR amplifying 71 non-coding DNA sequence intervals (see Methods; 43 sequences at *LMX1A* and 28 at *LMX1B*). The putative REs were cloned into pGWcfos: EGFP (Fisher et al. 2006a, b) and injected into fertilized zebrafish embryos. All zebrafish embryos showing mosaic EGFP reporter expression [33 *LMX1A* (77 %) and 14 *LMX1B* (50 %) elements] were separated to be raised for germline transmission analysis.

Assayed sequences display LMX1A and LMX1B-appropriate neuronal enhancer activities

Of the assayed sequences, 37 displayed reporter expression in the CNS upon passage through the germline. We identified two or more founders with concordant expression in

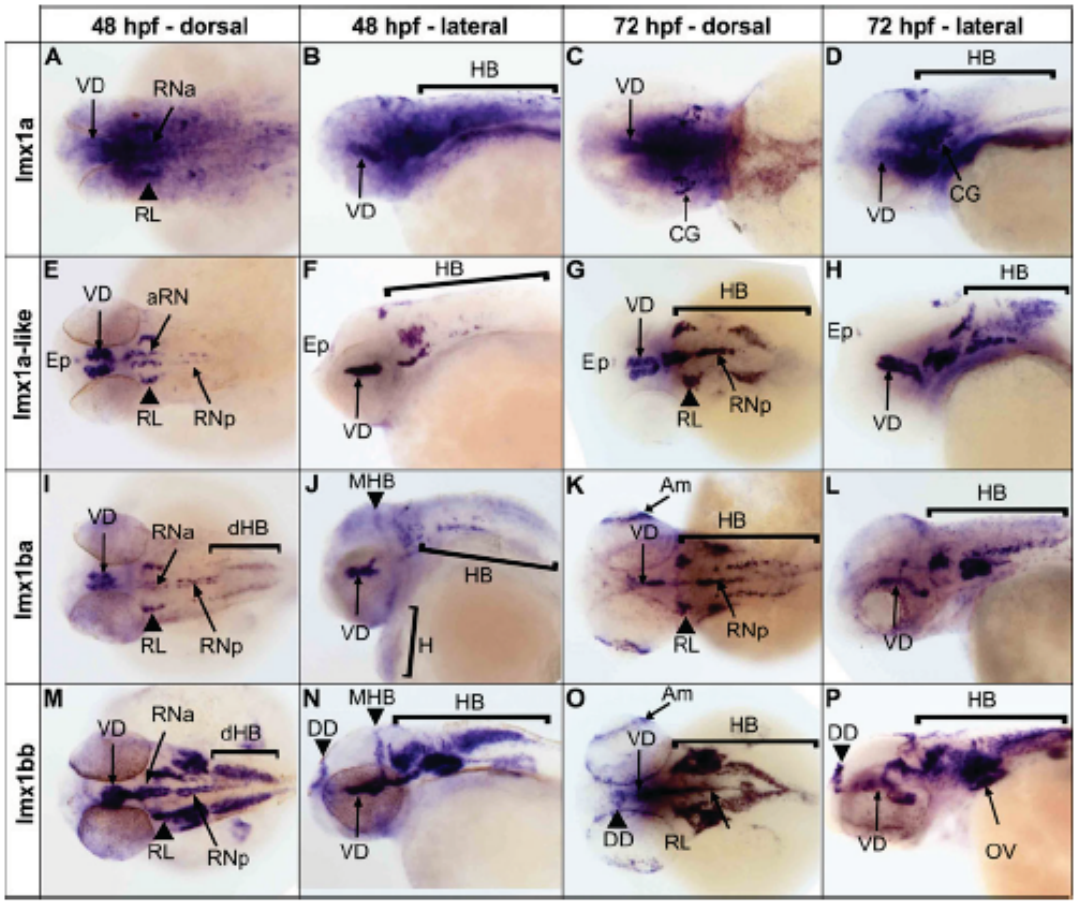


Fig. 1 In situ hybridization depicting the expression patterns of endogenous zebrafish *LMX1A* and *LMX1B* orthologs. Expression of *lmx1a* (**a**–**d**), *lmx1a-like* (**e**–**h**), *lmx1ba* (**i**–**l**) and *lmx1bb* (**m**–**p**) are shown, assayed at 48 hpf and 72 hpf. *Am* amacrine neurons, *CG* cranial ganglia, *DD* dorsal diencephalon, *dHB* dorsal hindbrain, *Ep* epiphysis, *H* heart, *Hb* hindbrain, *Mb–Hb* midbrain–hindbrain boundary, *OV* otic vesicle, *RL* rhombic lip, *RNa* anterior raphe nuclei, *RNp* posterior raphe nuclei, *VD* ventral diencephalon. Anterior is shown to the left in each panel

22/37 (59 %) of these sequences (*LMX1A*, $n = 17$ and *LMX1B*, $n = 5$). Those lines, for which we were unable to identify multiple founders, in general suffered from poor survival and fecundity. Those elements were therefore most often excluded, not because of divergent expression patterns but due to the inability to obtain a sufficient number of fertilized embryos. All 22 sequences displayed spatial control in the CNS in a manner consistent with aspects of *LMX1A* (Failli et al. 2002) and *LMX1B* (Chen et al. 2003) and with endogenous zebrafish patterns of expression described above. This includes directing reporter expression within discrete regions of the diencephalon, telencephalon and hindbrain. In addition, we identify enhancers at *LMX1A* that display regulatory control of reporter expression resembling the more diffuse expression of *lmx1a* in the CNS (*LMX1A*_ − 1.59). Consistent with their neuronal activity in our synthetic assay, the majority of endogenous sequence intervals corresponding to our assayed sequences display enrichment for histone 3 lysine 4 monomethylation (H3K4me1), a histone mark enriched at enhancers, in cultured neurospheres derived from human neuronal cells: NGED (neurosphere cultured cells, ganglionic eminence derived) and NCD (neurosphere cultured cells, cortex derived) (Fig. 2a–c) (Bernstein et al. 2010). Indeed, despite lacking a positive GERP alignment score, sequence *LMX1A*_36.74 displayed

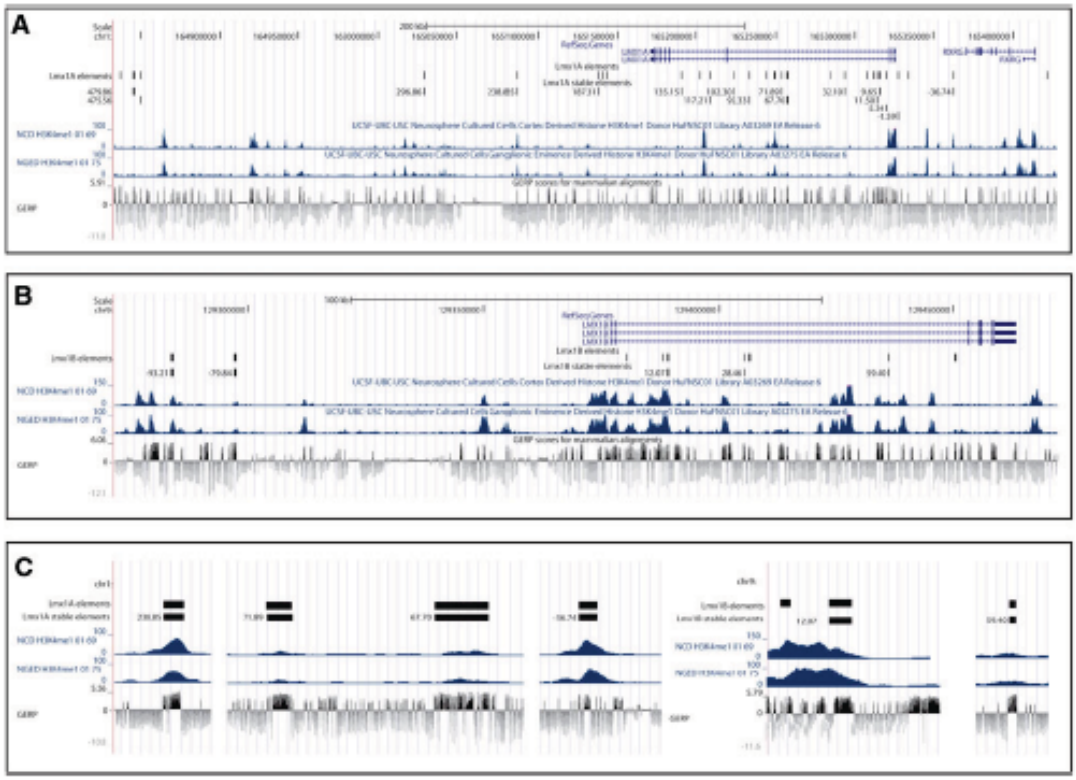


Fig. 2 *LMX1A* (a) and *LMX1B* (b) genomic loci displaying the selected sequences and their corresponding GERP sequence conservation tracks. H3K4me1 ChIP-seq signal is included from two types of cultured neurospheres, cortex derived and ganglionic eminence derived, showing substantial overlap between conservation and high H3K4me1 signal intensity. Panel C provides enlarged example intervals to indicate local sequence conservation within amplicons. The names of REs indicate approximate distance in kb from the start codon of the gene

strong H3K4me1 binding in both NGED and NCD cells and was also validated in our zebrafish assay.

Identification of LMX1A enhancers with telencephalic and diencephalic regulatory control

Consistent with the endogenous expression of the mouse *Lmx1a* mammalian ortholog (Failli et al. 2002), we identified many *LMX1A* enhancers displaying overlapping control in the telencephalon (Fig. 3a–e; Figure S2 and Table 1). Telencephalic expression displayed by these sequences is consistent with the function of LMX1 genes in cortical hem development (Fig. 3a–e) (Chizhikov et al. 2010; Adams et al. 2000; Guo et al. 2007). Telencephalic expression is also evident for the zebrafish *lmx1a*, although diffuse and not at significant levels (Fig. 1a–d). This observation may thus reflect mammalian (*LMX1A*/*Lmx1a*) control alone in this structure. In addition, we identify multiple sequences at *LMX1A* that direct expression in the diencephalon (Fig. 3a, d–f; Figure S2 and Table 1; e.g., *LMX1A*_238.85, *LMX1A*_ − 36.74 and *LMX1A*_9.65). These populations may include portions of the catecholaminergic diencephalic cluster, consistent with the established role of *Lmx1a* in mouse catecholaminergic neurogenesis (Yan et al. 2011).

Many identified LMX1A and LMX1B enhancers display regulatory control at the midbrain/hindbrain boundary and in the hindbrain

Multiple REs from both loci are able to drive expression in the midbrain–hindbrain boundary region that includes the IsO and anterior cerebellum (Fig. 4, Figure S2 and Table 1; e.g., *LMX1A*_ − 36.74, *LMX1A*_41.10, *LMX1A*_ 479.86, *LMX1A*_135.15, and *LMX1B*_ − 79.84). These data corroborate with the role of mammalian *LMX1* genes in IsO and cerebellum development and function (Adams et al. 2000; Guo et al. 2007). In addition, many assayed

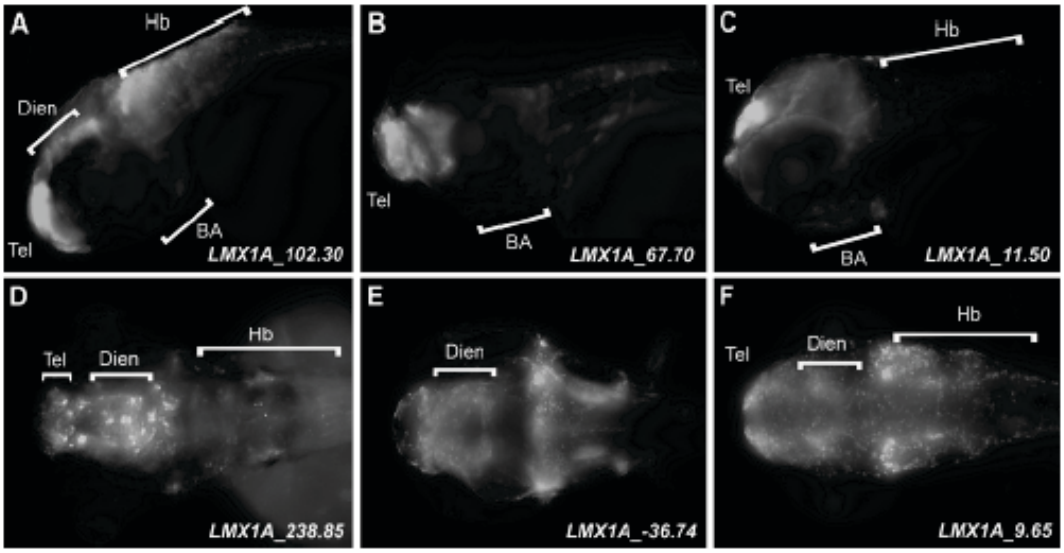


Fig. 3 Expression of EGFP reporter in the diencephalon and telencephalon of six representative transgenic zebrafish lines. Zebrafish embryos were fixed at 72 hpf and stained with anti-GFP antibody. Anterior is shown to the left in each panel. **a**–**c** Lateral images. **d**–**f** Dorsal images. *BA* branchial arches, *Hb* hindbrain, *Dien* diencephalon, *Hb* hindbrain, *Tel* telencephalon

sequences directed expression in the hindbrain (Fig. 4, Figure S2 and Table 1; e.g., *LMX1A*_475.57, *LMX1A*_238.85, *LMX1A*_5.34, *LMX1A*_ − 1.59, and *LMX1B*_28.46), including the roof plate (*LMX1A*_102.3) and the spinal cord (Figure S2 and Table 1; e.g., *LMX1A*_ − 1.59, *LMX1A*_238.85, and *LMX1B*_28.46) consistent with the endogenous expression of their corresponding zebrafish paralogous transcripts.

LMX1 enhancers display regulatory control in peripheral neuronal as well as non-neuronal cell populations

We identified several *LMX1A* and *LMX1B* sequences that direct expression in the otic vesicle (Fig. 3b, Figure S2 and Table 1; e.g., *LMX1A*_ − 36.74, *LMX1A*_92.33 and *LMX1A*_117.21, *LMX1B*_ − 93.21), consistent with the expression of *Lmx1a* (Failli et al. 2002). Furthermore, mice deficient in *Lmx1a* display abnormal ear development and deafness (Millonig et al. 2000; Huang et al. 2008). Multiple lines display reporter expression in PNS structures, such as motor neurons (Figure S2 and Table 1; *LMX1A*_238.85) or sympathetic chain (Figure S2 and Table 1; e.g. *LMX1A*_475.57). In contrast to the largely neuronal roles of *LMX1A* and *LMX1B*, many identified enhancers also drive reporter expression in non-neuronal tissues such as the branchial arches (Table 1, Fig. 3a–c, Figure S2), which correspond to documented mouse expression (Chen et al. 2003; Failli et al. 2002). One *LMX1B* enhancer also displays expression in the heart (*LMX1B*_ − 93.21) consistent with the endogenous expression of *lmx1ba*. The biological significance of this expression has not yet been determined but may, in part, correspond to the aortic arch neurons where expression of the catecholaminergic marker tyrosine hydroxylase has been previously reported (Wen et al. 2008).

Discussion

To better understand the regulatory landscape of *LMX1A* and *LMX1B*, we undertook a functional study of conserved, non-coding sequences (putative REs) at these loci, using zebrafish transgenesis. We first established the identity of two zebrafish paralogs for each human LMX1 gene. We then demonstrate that their endogenous expression closely resembles the previously characterized expression of their mammalian counterparts, including expression in the areas of presumptive catecholaminergic neurons, cerebellum, raphe nuclei and otic vesicles. Next, we used comparative sequence analyses to identify conserved, non-coding sequences at the human *LMX1A* and *LMX1B* loci successfully amplifying 71 putative REs for functional evaluation. Of these, 45 directed CNS reporter expression in mosaic zebrafish embryos. We further described the reporter expression of 22 sequences in stable transgenic

Table 1 Systematic annotation of *LMX1A* and *LMX1B* enhancers' activity in zebrafish body structures

Enhancer	Telen	Dien	Mesen	Rhomben	Sp Co	Mb/Hb	PNS	Other
*LMX1A*_ − 36.74	+	+	+	+		++	+	
*LMX1A*_ − 1.59	++	++	++	++	++	+	+	
*LMX1A*_5.34				+	++	+		NTC
*LMX1A*_9.65	+	+		+	+			OV
*LMX1A*_11.50	+	+	+	+		+	+	
*LMX1A*_32.09		+	+	+	+			Pn, L
*LMX1A*_67.70	++	+			+		+	NTC
*LMX1A*_71.89	+	+	+	+			+	C, F, H, R, OV
*LMX1A*_92.33	+				+			OV, C, F
*LMX1A*_102.30	++	+	+	++	+	+		
*LMX1A*_117.21	+	+	+	+	+	+	+	H, C
*LMX1A*_135.15	+	+	+	+	+	++	+	
*LMX1A*_187.31	+	+	+	+	+		++	
*LMX1A*_238.85		+	+	++	++		++	
*LMX1A*_296.85	++	+		++	+			
*LMX1A*_475.56	+	+	+	+	+		+	OV
*LMX1A*_479.56	+		+	+		++		C
*LMX1B*_ − 93.21				+			+	H, C
*LMX1B*_ − 79.84			++	+	+	++	+	
*LMX1B*_12.07	+	+	+	+	+	+	+	U
*LMX1B*_28.46	+	+	+	+	+	+		L
*LMX1B*_59.40	+	+	+	+	+		+	

Telen telecephalon, *Dien* diencephalon, *Mesen* mesencephalon, *Rhomben* rhombencephalon, *SpCo* spinal cord, *Mb/Hb* midbrain/hindbrain, *PNS* peripheral nervous system, *H* heart, *C* cartilage, *OV* otic vesicle, *L* lens, *Ub* ubiquitous, *NTC* notochord, *F* fins, *Pn* pronephros, *R* retina, *B* blood
+, weak; ++, moderate; +++, strong expression (relative determination)

lines (*LMX1A*, $n = 17$; *LMX1B*, $n = 5$). All 22 display consistent CNS enhancer function ($n \geq 2$ independent founders) that overlaps, at least in part, with the endogenous transcripts. The majority of these sequences display enrichment for H3K4me1, a modification known to be enriched at enhancers (Heintzman et al. 2007), in cultured neurospheres (Bernstein et al. 2010), consistent with their neuronal activity in our synthetic assay, and providing evidence supporting their likely *cis*-regulatory role in their endogenous context.

The diencephalon, telencephalon and midbrain–hindbrain boundary were among the most common structures marked by reporter expression for REs identified at both loci (Fig. 3, Table 1, Figure S2). Many enhancers similarly directed broad expression in the midbrain (Fig. 3, Table 1, Figure S2) and more discrete expression in the hindbrain e.g., in single rhombomeres, area postrema (Fig. 4, Table 1, Figure S2; *LMX1B*_−79.84) or hindbrain roof plate (Fig. 3, Table 1, Figure S2; *LMX1A*_102.3). These sites of expression overlap known domains of *Lmx1a* and *Lmx1b* expression in mammals and teleosts.

The expression directed in the midbrain–hindbrain boundary, cerebellum and posterior rhombomeres is consistent with the important function of both *LMX1A* and *LMX1B* in the development of the cerebellum rhombic lip and hindbrain roof plate (Mishima et al. 2009; Chizhikov et al. 2010). Furthermore, *LMX1B* is known to be instrumental for proper functioning of the IsO (Adams et al. 2000; Guo et al. 2007). We speculate that the forebrain expression of *LMX1A* enhancers might reflect its role in early development of cortical hem in mammals.

Many sequences also direct expression in the PNS and some non-neuronal tissues, consistent with the endogenous expression of these TFs and their critical role in the differentiation and maintenance of a range of populations. The most common non-neuronal sites of reporter expression are in the otic vesicles, which is consistent with *LMX1A/B* biology (Millonig et al. 2000; Failli et al. 2002; Huang et al. 2008). We also see some enhancers driving expression in the heart that may correspond to peripheral neuronal populations (Wen et al. 2008).

Importantly, transgenic assays provide an approximation of how a regulatory sequence can behave in a model system and may not capture every nuance that the corresponding sequence may display in context. Furthermore, their correspondence to spatial expression of *lmx1* genes

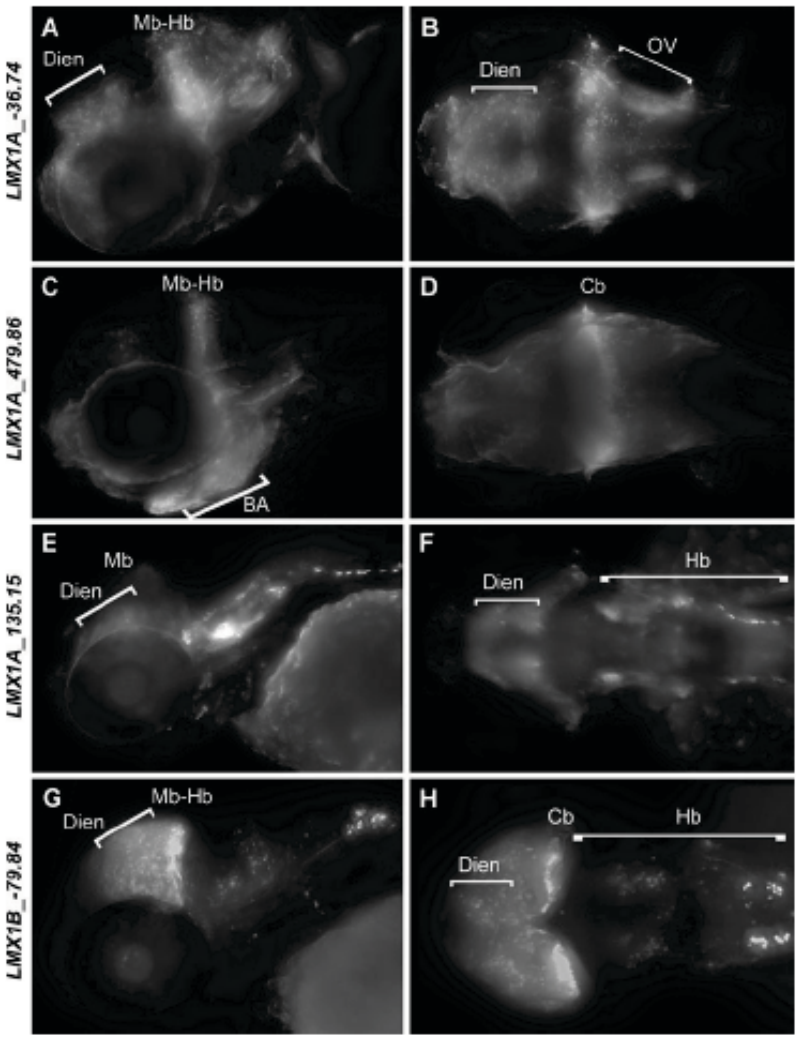


Fig. 4 Expression of EGFP reporter in the hindbrain and midbrain–hindbrain boundary of four representative transgenic zebrafish lines. Zebrafish embryos were fixed at 72 hpf and stained with anti-GFP antibody. Anterior is shown to the left in each panel. **a**, **c**, **e**, **g** are lateral images. **b**, **d**, **f**, **h** are dorsal images. *Cb* cerebellum, *Dien* diencephalon, *Hb* hindbrain, *Mb* midbrain, *Mb–Hb* midbrain–hindbrain boundary, *OV* otic vesicle

does not definitively demonstrate their control (exclusive or shared with neighbors) of these genes. In particular, we recognize that aspects of CNS regulatory control displayed by enhancers isolated at *LMX1A* and *LMX1B* may also be equally considered consistent with flanking gene expression. In particular, expression of the zebrafish *pbx1a* paralog includes many domains that also show expression of *lmx1a*/*lmx1a-like* (Thisse et al. 2004), including discrete expression in the telencephalon. Thus, firm conclusions regarding enhancer driven reporter expression and their direct relation to LMX1-expressing neuronal populations or those of their flanking genes will require additional experimental determination of possible physical interaction between enhancer and one or more cognate promoter. Despite these caveats, these assays can and do provide significant insight into the biological relevance of assayed sequences.

When it comes to genome annotation, there is no satisfactory "one size fits all" approach. We demonstrate how a range of available data types may be integrated in the exploration of the genomic information content of sequence encompassing two critical human genes. This work describes the endogenous expression patterns of zebrafish *LMX1A* paralogs, identifies 22 previously unknown enhancers and sheds light on previously unknown transcriptional regulatory landscape at the *LMX1A* and *LMX1B* loci. If one accounts for the presence of additional conserved and/or histone-marked sequences

in the genomic intervals under consideration, these enhancers may represent only a fraction of the conserved non-coding elements at these loci. We hypothesize that many enhancers may be required in combination to orchestrate regulatory control of these genes. The pleiotropy of neuronal subsets marked by these identified enhancers may highlight additional complexity in this regulatory control or reflect position effects. These data reinforce the value of targeted screens in the analysis of human disease loci integrating comparative sequence analysis, chromatin modifications and functional validation using zebrafish transgenesis in the identification of transcriptional regulatory sequences.

Methods

Fish maintenance

Zebrafish were kept and bred under standard conditions at 28.5 °C (Westerfield 2000). Embryos were staged and fixed at 48 and 72 hpf using 4 % paraformaldehyde (PFA) in phosphate-buffered saline (PBS; pH 7.2) as described elsewhere (Kimmel et al. 1995). To better visualize in situ hybridization and *EGFP* reporter results, embryos were grown in 0.2 mM 1-phenyl-2-thiourea (Sigma) to inhibit pigment formation (Westerfield 2000).

Whole mount in situ hybridization

Digoxigenin labeled riboprobes complementary to *lmx1a*, *lmx1a-like*, *lmx1ba* or *lmx1bb* mRNAs were generated by linearization of pCR II TOPO TA vectors containing partial ORFs of the genes (for probe sequences see Figure S3). Plasmids were linearized with *Eco*RV (New England Biolabs) and subsequently labeled riboprobes were transcribed using SP6 polymerase and the DIG RNA Labeling Kit (T7/SP6) (Roche). Probes were synthesized for 2 h at 37 °C, followed by the addition of 1 µl of RNAse free DNAse I for DNA template digestion. Subsequently, probes were purified using SigmaSpin columns (Sigma–Aldrich). Whole mount in situ hybridization reactions were performed using 1:4,000 dilutions of rioboprobes at 70 °C as previously described (Thisse et al. 1993, 2003)—see http://zfin.org/zf_info/zfbook/chapt9/9.82.html for detailed protocol. Probe sequences were selected to avoid cross-hybridization with *lmx1* family members and unrelated transcripts using pairwise alignment of *lmx1* transcripts to find unique stretches of mRNA. Sequences were aligned using Clustal Omega (http://www.ebi.ac.uk/Tools/msa/clustalo/). All probe sequences and corresponding oligonucleotides for their amplification are provided in supplemental data (Figure S4 and S5).

Selection and amplification of human non-coding sequences

To select regions to test for potential enhancer activity, genomic intervals encompassing *LMX1A* and *LMX1B* loci were considered up to the neighboring genes, set as boundaries (*LMX1A*, chr1: 163,082,934–163,636,974 bp; *LMX1B*, chr9: 128,309,140–128,607,182 bp). This study is not intended to be exhaustive. The genomic intervals encompassing these genes are very large. Thus, sequences were prioritized for selection based upon proximity to the *LMX* loci and conservation, but no ranking of conservation was applied. Consequently, this approach is likely to have identified only a subset of enhancers regulating *LMX1A* and *LMX1B*. Using Galaxy computational interface (Goecks et al. 2010) and UCSC genome browser, we chose conserved non-coding vertebrate elements with positive GERP alignment scores (Cooper et al. 2005). The GERP algorithm identifies constrained sequences in genomic alignments by determining whether a paucity of substitutions exists at each point in an alignment compared to what one expects of the neutral rate of evolution. Selected intervals positioned <500 bp apart were merged into single amplicons. DNA region coordinates and primer sequences used for amplification are listed in table S2. Sequences were also selected to avoid clustering and are distributed across these loci (Fig. 1a). Amplicons were PCR amplified from human genomic DNA, TA cloned into pCR8 (Invitrogen, USA) and then cloned using the Gateway system (Invitrogen, USA) into pGW_*cfos*EGFP as previously described (Fisher et al. 2006a, b).

Embryo injection and analysis

EGFP reporter constructs were injected into AB background G0 embryos ($n \geq 200$) at the one-to two-cell stage with tol2 transposase as previously described (Fisher et al. 2006a, b). Injected embryos were evaluated for EGFP expression between 24 and 72 hpf. As negative controls, *EGFP* reporter constructs containing only the cfos promoter were injected. Nonspecific expression from the cfos minimal promoter is occasionally observed in the myotome and no other nonspecific expression was detected (data not shown). Embryos showing consistent *EGFP* expression were selected and raised for further analysis when signal was observed in ≥10 % of injected embryos. Mosaic fish were subsequently crossed to identify those constructs that passed through the germline transmission, better facilitating spatial evaluation of corresponding *EGFP* expression. Instances where we do not report expression reflect failure to identify more than one founder transgenic line and not inconsistencies among multiple lines for a single construct. Re-injection and additional screening may help resolve the

neuronal regulatory control of additional constructs we have generated at these loci but whose activities we do not report upon here. Embryos were imaged using a Carl Zeiss Lumar V12 Stereo microscope with AxioVision software (version 4.5).

Immunocytochemistry

Embryos were anesthetized with tricaine (10 μg/ml) in embryo medium (Westerfield 2000) and fixed in 4 % PFA in phosphate-buffered saline (PBS; pH 7.2) for 2 h. They were then rinsed four times in PBST (PBS/0.1 % Triton X-100), incubated in Proteinase K for 1 h at room temperature, washed 5 × 5 min in PBST, and incubated for 2 h in blocking solution [10 % goat serum, 1 % bovine serum albumin (BSA), in PBST]. Embryos were then incubated overnight at room temperature in primary antibody (anti-GFP, Invitrogen 1:2,000), rinsed 6 × 45 min in PBST 1 % goat serum, and incubated overnight at room temperature in secondary antibody (Alexa-Fluor, 488, Invitrogen 1:1,000). They were then rinsed 5 × 10 min in PBST and transferred to 50 % glycerol in PBS for imaging.

Acknowledgments The authors gratefully acknowledge the support of the McKusick Nathans Institute of Genetic Medicine Center for Functional Investigation in Zebrafish (FINZ). This research was supported in part by the National Institute of Neurological Disease and Stroke (R01 NS062972; NINDS, NIH) to ASM. XR was also supported by NIH pre-doctoral training grant 5T32GM07814. SM was supported by funds from the National Institute of Standards and Technology.

Certain commercial equipment, instruments, materials, or companies are identified in this paper to specify adequately the experimental procedure. Such identification does not imply recommendation nor endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are the best available for the purpose. *Official contribution of NIST; not subject to copyright.

Ethical standard All zebrafish work was performed under an approved protocol (FI10M369), reviewed by the Johns Hopkins Institutional Animal Care and Use Committee.

References

Adams KA, Maida JM, Golden JA, Riddle RD (2000) The transcription factor Lmx1b maintains Wnt1 expression within the isthmic organizer. Development 127:1857–1867

Amores A, Force A, Yan YL, Joly L, Amemiya C et al (1998) Zebrafish hox clusters and vertebrate genome evolution. Science 282:1711–1714

Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A et al (2010) The NIH roadmap epigenomics mapping consortium. Nat Biotechnol 28:1045–1048

Chen X, Mao T, Dai Y (2003) Experimental study of artificial bone composite of bicoral, rhBMP-2 and PLA in repairing calvarial defects. Hua Xi Kou Qiang Yi Xue Za Zhi 21:474–476

Cheng CW, Yan CH, Choy SW, Hui MN, Hui CC, Cheng SH (2007) Zebrafish homologue irx1a is required for the differentiation of serotonergic neurons. Dev Dyn 236(9):2661–2667

Chizhikov VV, Lindgren AG, Mishima Y, Roberts RW, Aldinger KA et al (2010) Lmx1a regulates fates and location of cells originating from the cerebellar rhombic lip and telencephalic cortical hem. Proc Natl Acad Sci USA 107:10725–10730

Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S et al (2005) Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 15:901–913

Dai JX, Johnson RL, Ding YQ (2009) Manifold functions of the nail-patella syndrome gene Lmx1b in vertebrate development. Dev Growth Differ 51:241–250

Ding YQ, Marklund U, Yuan W, Yin J, Wegman L et al (2003) Lmx1b is essential for the development of serotonergic neurons. Nat Neurosci 6:933–938

Elsen GE, Choi LY, Millen KJ, Grinblat Y, Prince VE (2008) Zic1 and Zic4 regulate zebrafish roof plate specification and hindbrain ventricle morphogenesis. Dev. Biol 314(2):376–392

Failli V, Bachy I, Retaux S (2002) Expression of the LIM-homeodomain gene Lmx1a (dreher) during development of the mouse nervous system. Mech Dev 118:225–228

Filippi A, Mahler J, Schweitzer J, Driever W (2010) Expression of the paralogous tyrosine hydroxylase encoding genes th1 and th2 reveals the full complement of dopaminergic and noradrenergic neurons in zebrafish larval and juvenile brain. J Comp Neurol 518:423–438

Fisher S, Grice EA, Vinton RM, Bessling SL, McCallion AS (2006a) Conservation of RET regulatory function from human to zebrafish without sequence similarity. Science 312:276–279

Fisher S, Grice EA, Vinton RM, Bessling SL, Urasaki A et al (2006b) Evaluating the biological relevance of putative enhancers using Tol2 transposon-mediated transgenesis in zebrafish. Nat Protoc 1:1297–1305

Goecks J, Nekrutenko A, Taylor J (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 11:R86

Guo C, Qiu HY, Huang Y, Chen H, Yang RQ et al (2007) Lmx1b is essential for Fgf8 and Wnt1 expression in the isthmic organizer during tectum and cerebellum development in mice. Development 134:317–325

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW et al (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 39:311–318

Hobert O, Westphal H (2000) Functions of LIM-homeobox genes. Trends Genet 16:75–83

Huang M, Sage C, Li H, Xiang M, Heller S et al (2008) Diverse expression patterns of LIM-homeodomain transcription factors (LIM-HDs) in mammalian inner ear development. Dev Dyn 237:3305–3312

Kimmel CB, Ballard WW, Kimmel SR, Ullmann B, Schilling TF (1995) Stages of embryonic development of the zebrafish. Dev Dyn 203:253–310

Kleinjan DJ, Coutinho P (2009) *Cis*-ruption mechanisms: disruption of *cis*-regulatory control as a cause of human genetic disease. Brief Funct Genomic Proteomic 8:317–332

Koo SK, Hill JK, Hwang CH, Lin ZS, Millen KJ et al (2009) Lmx1a maintains proper neurogenic, sensory, and non-sensory domains in the mammalian inner ear. Dev Biol 333:14–25

Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E et al (2012) Systematic localization of common disease-associated variation in regulatory DNA. Science 337:1190–1195

McMahon C, Gestri G, Wilson SW, Link BA (2009) Lmx1b is essential for survival of periocular mesenchymal cells and influences Fgf-mediated retinal patterning in zebrafish. Dev Biol 332:287–298

Millonig JH, Millen KJ, Hatten ME (2000) The mouse Dreher gene Lmx1a controls formation of the roof plate in the vertebrate CNS. Nature 403:764–769

Mishima Y, Lindgren AG, Chizhikov VV, Johnson RL, Millen KJ (2009) Overlapping function of Lmx1a and Lmx1b in anterior hindbrain roof plate formation and cerebellar growth. J Neurosci 29:11377–11384

Nakatani T, Kumai M, Mizuhara E, Minaki Y, Ono Y (2010) Lmx1a and Lmx1b cooperate with Foxa2 to coordinate the specification of dopaminergic neurons and control of floor plate cell differentiation in the developing mesencephalon. Dev Biol 339:101–113

Nichols DH, Pauley S, Jahan I, Beisel KW, Millen KJ et al (2008) Lmx1a is required for segregation of sensory epithelia and normal ear histogenesis and morphogenesis. Cell Tissue Res 334:339–358

Noonan JP, McCallion AS (2011) Genomics of long-range regulatory elements. Annu Rev Genomics Hum Genet 11:1–23

O'Hara FP, Beck E, Barr LK, Wong LL, Kessler DS, Riddle RD (2005) Zebrafish Lmx1b.1 and Lmx1b.2 are required for maintenance of the isthmic organizer. Development 132(14): 3163–3173

Shirasaki R, Pfaff SL (2002) Transcriptional codes and the control of neuronal identity. Annu Rev Neurosci 25:251–281

Thisse C, Thisse B, Schilling TF, Postlethwait JH (1993) Structure of the zebrafish snail1 gene and its expression in wild-type, spadetail and no tail mutant embryos. Development 119:1203–1215

Thisse C, Degrave A, Kryukov GV, Gladyshev VN, Obrecht-Pflumio S et al (2003) Spatial and temporal expression patterns of selenoprotein genes during embryogenesis in zebrafish. Gene Expr Patterns 3:525–532

Thisse B, Heyer V, Lux A, Alunni A, Degrave A, Seiliez I, Kirchner J, Parkhill J-P, Thisse C (2004) Spatial and temporal expression of the zebrafish genome by large-scale in situ hybridization screening. Meth Cell Biol 77:505–519

Visel A, Rubin EM, Pennacchio LA (2009) Genomic views of distant-acting enhancers. Nature 461:199–205

Vollrath D, Jaramillo-Babb VL, Clough MV, McIntosh I, Scott KM et al (1998) Loss-of-function mutations in the LIM-homeodomain gene, LMX1B, in nail-patella syndrome. Hum Mol Genet 7:1091–1098

Wen L, Wei W, Gu W, Huang P, Ren X, Zhang Z, Zhu Z, Lin S, Zhang B (2008) Visualization of monoaminergic neurons and neurotoxicity of MPTP in live transgenic zebrafish. Dev. Biol 314(1):84–92

Westerfield M (2000) The zebrafish book: a guide for the laboratory use of zebrafish (Danio rerio). University of Oregon Press, Eugene

Yan CH, Levesque M, Claxton S, Johnson RL, Ang SL (2011) Lmx1a and lmx1b function cooperatively to regulate proliferation, specification, and differentiation of midbrain dopaminergic progenitors. J Neurosci 31:12413–12425

This is a License Agreement between Xylena Reed ("You") and Springer ("Springer") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Springer, and the payment terms and conditions.

License Number	3553090399818
License date	Jan 20, 2015
Licensed content publisher	Springer
Licensed content publication	Molecular Genetics and Genomics
Licensed content title	Integration of genomic and functional approaches reveals enhancers at LMX1A and LMX1B
Licensed content author	Grzegorz M. Burzynski
Licensed content date	Jan 1, 2013
Volume number	288
Issue number	11
Type of Use	Thesis/Dissertation
Portion	Full text
Number of copies	3
Author of this Springer article	Yes and you are the sole author of the new work
Order reference number	None
Title of your thesis / dissertation	Towards the genesis of neuronal regulatory catalogs and their vocabularies
Expected completion date	Jan 2015

Springer and the original publisher: Molecular Genetics and Genomics, Volume 288, Nov 2013, pp 579-589; Integration of genomic and functional approaches reveals enhancers at LMX1A and LMX1B, Grzegorz M. Burzynski, Xylena Reed, Samantha Maragh, Takeshi Matsui, Andre S. McCallion; figures 1-4, © Springer-Verlag Berlin Heidelberg 2013; with kind permission from Springer Science and Business Media.

Appendix 4

Method

Systematic elucidation and in vivo validation of sequences enriched in hindbrain transcriptional control

Grzegorz M. Burzynski,[1,4] Xylena Reed,[1,2,4] Leila Taher,[3,4] Zachary E. Stine,[1] Takeshi Matsui,[1] Ivan Ovcharenko,[3,5] and Andrew S. McCallion[1,5]

[1]*McKusick–Nathans Institute of Genetic Medicine, Department of Molecular and Comparative Pathobiology, The Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA;* [2]*Predoctoral Training Program in Human Genetics, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA;* [3]*Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA*

Illuminating the primary sequence encryption of enhancers is central to understanding the regulatory architecture of genomes. We have developed a machine learning approach to decipher motif patterns of hindbrain enhancers and identify 40,000 sequences in the human genome that we predict display regulatory control that includes the hindbrain. Consistent with their roles in hindbrain patterning, MEIS1, NKX6-1, as well as HOX and POU family binding motifs contributed strongly to this enhancer model. Predicted hindbrain enhancers are overrepresented at genes expressed in hindbrain and associated with nervous system development, and primarily reside in the areas of open chromatin. In addition, 77 (0.2%) of these predictions are identified as hindbrain enhancers on the VISTA Enhancer Browser, and 26,000 (60%) overlap enhancer marks (H3K4me1 or H3K27ac). To validate these putative hindbrain enhancers, we selected 55 elements distributed throughout our predictions and six low scoring controls for evaluation in a zebrafish transgenic assay. When assayed in mosaic transgenic embryos, 51/55 elements directed expression in the central nervous system. Furthermore, 30/34 (88%) predicted enhancers analyzed in stable zebrafish transgenic lines directed expression in the larval zebrafish hindbrain. Subsequent analysis of sequence fragments selected based upon motif clustering further confirmed the critical role of the motifs contributing to the classifier. Our results demonstrate the existence of a primary sequence code characteristic to hindbrain enhancers. This code can be accurately extracted using machine-learning approaches and applied successfully for de novo identification of hindbrain enhancers. This study represents a critical step toward the dissection of regulatory control in specific neuronal subtypes.

[Supplemental material is available for this article.]

In metazoans, precise spatiotemporal patterns of gene expression are modulated by the exquisite contributions of transcriptional regulatory sequences. These include enhancers that activate transcription in a manner frequently observed to be independent of distance, position, and orientation with respect to the promoter of their target genes (Banerji et al. 1981). Empirically validated enhancers are typically a few hundred base pairs long and comprise binding sites for multiple transcription factors (TFs). In turn, TFs bound to these sequences also interact with common co-activators, communicating with the basal transcription machinery assembled at the promoter, and increasing the rate of transcription (Bulger and Groudine 2011). Identifying the combinatorial protein–DNA and protein–protein interactions that determine spatial and temporal enhancer function is crucial to understanding how distinct cellular and developmental programs are established.

The systematic discovery of enhancers has proven challenging, since they are often located at great genomic distances from the genes they regulate (Letticeet al. 2003). The classical approach to enhancer identification involves the use of sequence constraint in the proximity to genes with known biology or expression in a tissue of interest. However, this approach is limited in that comparative genomics offers no information regarding the specific regulatory role of the sequences (Noonan and McCallion 2010). Recent advances in sequencing technologies have enabled the identification of protein–DNA interactions and chromatin structural conformation at the whole-genome level (Barski and Zhao 2009; Visel et al. 2009; Ernst et al. 2011). For instance, the ENCODE project has annotated ~15 histone variants and modifications, as well as binding events for ~150 TFs and transcriptional co-factors in many human cell lines, identifying hundreds of thousands of sequence intervals harboring active chromatin (The ENCODE Project Consortium 2007). Despite the unprecedented scale of the ENCODE project, enhancers identified using the TFs, co-factors, and histone marks likely account for only a fraction of all tissue-specific enhancers utilized in any vertebrate (He et al. 2011). Identified sequences are tissue-specific and cannot be used to infer the gene regulatory activity in other tissues (Visel et al. 2009). The complete discovery and validation of enhancers in the human genome spanning all cell types and developmental stages will remain an elusive goal for years to come. Experimental efforts must be accompanied by large-scale computational predictions that are capable of deciphering the DNA sequence encoding tissue-specific regulatory elements and can be applied to annotate complete

[4]**These authors contributed equally to this work.**
[5]**Corresponding authors**
E-mail andy@jhmi.edu
E-mail ovcharen@nig.gov
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.139717.112. Freely available online through the *Genome Research* Open Access option.

22:2278–2289 © 2012, Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/12; www.genome.org

genomes. Accurate computational predictions not only permit whole-genome annotations of tissue-specific enhancers in a single species, but they can also be applied to annotation of related species in a straightforward manner (Lee et al. 2011). Computational approaches based on the analysis of sequence motifs shared among enhancers with the same or similar regulatory activities are not only capable of accurately predicting enhancers with specific biological functions de novo, but also contribute to our understanding of the combinatorial networks of TFs underlying particular spatio-temporal patterns of gene expression.

We previously proposed a novel computational strategy that combines comparative genomics, Gibbs sampling, and linear regression to systematically identify heart enhancers in the human genome (Narlikar et al. 2010). The reliability of our approach has been evaluated not only computationally, but also in vivo, using transgenic reporter assays in zebrafish and mouse, with a validation rate of 62% for our heart enhancer predictions. High-throughput experimental approaches, such as genome-wide chromatin immunoprecipitation coupled with massively parallel sequencing (ChIP-seq) against EP300, a histone acetyltransferase and transcriptional co-activator protein, predict the genomic location of heart developmental enhancers with comparable accuracy (Blow et al. 2010). These different strategies uncover only partially overlapping sets of putative heart enhancers. Thus, we observed only 17% of the sequences predicted by ChIP-seq experiments overlapping with our candidate heart enhancer sequences (Narlikar et al. 2010). Weak sequence conservation (Blow et al. 2010) alone does not explain this result, since ~80% of predictions based on ChIP-seq are conserved between human and mouse, which is the only evolutionary constraint imposed by our method. Instead, as current evidence suggests, the small overlap is more likely to be attributable to the different nature of the enhancer signatures captured by each model (He et al. 2011).

In this study we asked whether regulatory signatures (vocabularies) could be uncovered from a more complex cellular substrate, the central nervous system (CNS). In particular, we set out to determine the sequence basis of regulatory control in the hindbrain (Hb). The Hb, or rhombencephalon, is the most primitive part of the human brain, and likely evolved from a homologous structure present in Urbilateria around 550 million years ago (Ghysen 2003). It includes the cerebellum, pons, and medulla oblongata, which are structures that control functions as fundamental and diverse as respiration, heart rate, reflex, and voluntary movements. Impaired Hb development and function are associated with many disorders such as autism, ADHD (attention deficit hyperactivity disorder), schizophrenia, cerebral palsy, and various sleep disorders (Berquin et al. 1998; Aston-Jones 2005; Andreasen and Pierson 2008). As with other complex diseases and phenotypes, most variants identified by genome-wide association and sequencing population studies are found in noncoding regions of the genome, and therefore suspected to play a role in regulatory control (Cooper and Shendure 2011). Understanding the gene regulatory landscape of the human genome in Hb development and structure is an important step toward uncovering the noncoding substrate of the genomic component of brain disorders.

We introduce a machine learning approach, based on the distribution of transcription factor binding sites (TFBSs) in enhancers, which are capable of accurately identifying enhancers whose regulatory control includes the nascent Hb. Our classifier performs very well in de novo discovery of Hb enhancers, with 88% (30/34) of computational predictions validated in vivo using transgenic zebrafish reporter assays. We also analyze the impact of small collections TFBSs on the Hb function of the host enhancers, and present a map of 40,000 Hb enhancers in the human genome. In summary, our data show how the application of effective computational methods for enhancer prediction can greatly improve our understanding of the gene regulatory networks controlling human development and disease.

Results

Building a training set of Hb enhancers

In order to construct a model for Hb enhancer activity, we first compiled a data set of 211 enhancers for which Hb activity has been validated in vivo with reporter assay systems in transgenic mice and zebrafish (Supplemental Table S1). Most of these sequences ($n = 192$) were obtained from the VISTA Enhancer Browser (Visel et al. 2007) and an additional 20 enhancers were identified in our laboratory in the context of ongoing in vivo transgenic enhancer screens in zebrafish. This data set of Hb enhancers bears genomic features consistent with other enhancer sets. The GC and repeat-content of the Hb enhancers are close to the genome averages (Supplemental Fig. S1). Thirty-nine percent of the Hb enhancers in this catalog are intronic and 61% are intergenic, displaying a genomic distribution close to the expected (for comparison, 44% of enhancers in the VISTA database are intronic). On the other hand, these Hb enhancers are especially well conserved among vertebrates—99% of the Hb enhancers are conserved between human and mouse genomes, and 82% are also conserved between human and chicken genomes. The average phastCons evolutionary conservation score (Siepel et al. 2005) of Hb enhancers is 1.6, significantly higher than the corresponding scores of the heart and limb enhancers (0.5 and 1.2, respectively; Wilcoxon rank-sum test P-value << 0.001).

Enhancers driving expression in the nervous system frequently direct expression in one or more additional tissues or developmental stages. Eighty percent of Hb VISTA enhancers also direct transcription in other tissues, such as midbrain (49%), forebrain (33%), neural tube (43%), and limb (8%), suggesting that the same elements may play pleiotropic roles in expression, and thus that regulatory lexicons may not always be discrete.

Designing an enhancer classifier

There is now broad interest in determining the extent to which computational power can be used to elucidate how transcriptional regulatory instructions are encrypted in primary DNA sequence. The increased volume of genomic sequence-based data sets far exceeds our present capacity to impute biological value to primary sequence and variation therein, particularly in noncoding sequence. We previously developed a linear regression approach that relies on sequence patterns to accurately predict sequences with similar regulatory activity in the human genome de novo beginning with a small catalog of known heart enhancers (Narlikar et al. 2010). Since then, a similar method based on support vector machines (SVMs) and primitive short sequence segments (k-mers) has also performed well in classifying enhancers from different expression domains, including forebrain- and midbrain-derived ChIP-seq data sets (Lee et al. 2011). However, the SVM method was unable to accurately distinguish between different brain enhancer data sets. This was likely complicated in part by the vastly increased cellular complexity of the sequence used in their training sets. Therefore, although both the SVM and the linear regression method exhibited similar performances (data not shown), we

opted to combine the specificity of our original classifier with the advanced statistical model proposed by the latter approach (Lee et al. 2011). To this end, we constructed an SVM classifier operating on known TFBSs and overrepresented de novo identified motifs, which we dubbed EnhSVM (see Methods for details). We then used this strategy to determine if we could better discriminate among regulatory catalogs of CNS subdomains and extend this to define a classifier for the Hb, which currently has no ChIP-seq substrate available.

When applied to the collection of 11 tissue-specific experimentally validated sets of VISTA enhancers (forebrain, midbrain, hindbrain, neural tube, limb, heart, dorsal root ganglia, branchial arch, nose, cranial nerve, eye) our classifier was able to discriminate all enhancer sets from background genomic regions with accuracies exceeding 60% according to the area under the Receiver Operating Characteristic (ROC) curve (AUC) measurements in all cases (Supplemental Fig. S2). The vast majority of predictions produced by these models only overlapped predictions from related tissues, indicating that our method identifies cell type-specific enhancer signatures. CNS enhancer classifiers (forebrain, midbrain, hindbrain, neural tube) performed better than the rest (Supplemental Fig. S2), and the Hb classifier displayed the highest AUC accuracy at 91%.

Refinement of a hindbrain classifier

The embryonic Hb forms along the anterior–posterior axis and is initially segmented into a series of adjacent units called rhombomeres. The identity of these rhombomeres is correlated with domains of Hox gene expression and function, which in turn are determined by a gradient of retinoic acid along the anterior–posterior axis of the Hb (Schneider-Maunoury et al. 1998). Thus, the most anterior rhombomeres contribute to the metencephalon (pons and cerebellum), while the most posterior rhombomeres form the myelencephalon (medulla oblongata). In order to determine if we could further refine our classifier's predictive capacity, we separated the data set of Hb enhancers into 161 anterior and 153 posterior Hb enhancers based on expression patterns driven by the sequences in embryonic mice at developmental stage E11.5. The purpose of this step was twofold. First, although these two sets of enhancers are highly overlapping with ~80% of the sequences driving reporter expression in both domains, we hypothesized that simple functional clustering should result in increasingly homogeneous data sets, more suitable for our method. Second, combinations of multiple classifiers, in this case trained on different Hb subsets, often outperform single classifiers (Kittler et al. 1998). Consequently, we trained and tested three independent Hb classifiers using a standard 10-fold cross-validation setup on five random partitions of the data, using three slightly different data sets: the complete Hb data set, the subset of Hb enhancers that are active in the anterior Hb, and the subset of enhancers which functions in the posterior Hb. However, no single classifier significantly outperforms the others. Indeed, all three Hb classifiers achieved average AUCs of ~90%, with a true positive rate (TPR) of at least 47% at a false positive rate (FPR) of 5% (Fig. 1A).

Hindbrain enhancers harbor putative binding sites for transcriptional regulators of cell identity

Our Hb classifiers rely on sequence motifs representing TFBSs that facilitate distinction of Hb enhancers from random genomic sequences (Methods). We analyzed the discriminatory power of individual motifs to reveal specific TFs likely to interact with Hb enhancers. All three Hb classifiers identified motifs that are known to bind the critical Hb TFs MEIS1, NKX6-1, HOX family members, and POU protein family members among the 100 most relevant sequence features for identifying Hb enhancers (Waskiewicz et al. 2001; Nelson et al. 2006; Kiyota et al. 2008). Similarly, binding motifs known to bind SOX2, a TF which is highly expressed in the Hb with roles in CNS development, were common to all three Hb classifiers (Supplemental Table S2; Kelberman et al. 2008). Many of these motifs are specific to Hb development and function, and their relevance differs for analogous classifiers trained on data sets of enhancers specific to other tissues (compared, for example, with motifs relevant to limb and heart gene expression regulation, Supplemental Table S2; Supplemental Fig. S3). As expected, distinct sets of Hb sequences, even if largely overlapping, showed slight differences in the contribution of each motif to the decision function of the corresponding classifier. For example, we observed differences in the relevance of the estrogen receptor ESR1 motif, which is particularly enriched among enhancers active in the posterior Hb. Thus, the motif for ESR1 is among the 100 most relevant sequence features for the Hb classifier focusing on posterior Hb, but not among the 100 most relevant sequence features for the other two Hb classifiers. Estrogen receptor-related proteins, which can bind ESR1-like motifs (Vanacker et al. 1999; Giguere 2002), have previously been implicated in anterior–posterior brain

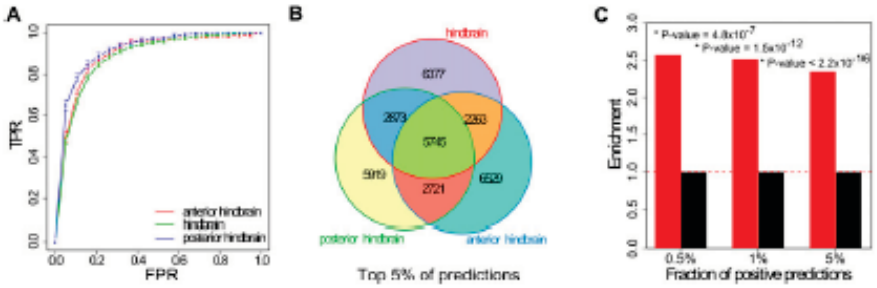


Figure 1. Hindbrain enhancers can be accurately predicted from DNA sequence. (*A*) Area under the ROC curve (AUC) for three Hb enhancer classifiers trained on three highly overlapping data sets (enhancers with activity in the anterior Hb, posterior Hb, and whole Hb). AUC values range from 0.5 (random discrimination) to a theoretical maximum of 1. We tested the performance of the classifiers in a cross-validation setting and obtained values of 0.89 (anterior Hb), 0.92 (posterior Hb), and 0.89 (combined Hb). (*B*) Overlap among the top-scoring 5% Hb enhancer predictions produced by all three Hb classifiers. (*C*) Fold-enrichment in 787 genes involved in Hb function in the neighborhood of positive predictions or putative Hb enhancers. Putative Hb enhancers were associated with the closest gene. *P*-values were computed using Fisher's exact test.

segmentation (Bardet et al. 2005). The ability of the Hb classifiers to recover motifs corresponding to known Hb TFs provides additional validation of our model. However, we must caution that it is likely that not all computationally predicted motifs are bound by a TF. Moreover, even if they are, assigning the identity of the TFs binding to these motifs is not straightforward, since the binding affinity catalog of TFs is not complete and many motifs are recognized by multiple TFs.

In order to determine the specificity of the motifs with high discriminatory power in the Hb classifiers, we compared them with those in forebrain, midbrain, and limb enhancer classifiers. These comparison classifiers were trained using EnhSVM on sequences identified using ChIP-seq with the enhancer-associated protein EP300 (Methods). While a negligible fraction (<5%) of EP300 peaks is shared among all data sets, overlap among EP300 peaks for closely related tissues, such as forebrain and midbrain, was higher (15%–20%), consistent with tissue-dependent EP300 binding specificity. Less than 10% of the motifs are shared among the 50 most relevant sequence features for the different classifiers. Additionally, <20% overlap with the motifs identified for Hb enhancers—an observation that highlights the ability of our Hb classifier to specifically capture the Hb enhancer code. The TFBSs shared by the Hb and other brain classifiers included binding sites for MEIS1, the NKX, SOX, and HOX homeobox factors, and ZHX2—developmental TFs that are characteristic of general brain regulatory pathways.

Genome-wide predictions identify novel hindbrain enhancers

Our training set is largely made up of conserved sequences and brain enhancers have been shown to frequently be deeply conserved (Visel et al. 2009), so to obtain a genome-wide map of putative human enhancers active in the Hb, we restricted our genome scan to sequences which are conserved among mammals, $n = 337{,}000$ (Siepel et al. 2005; Methods). We repeated the scans using the anterior Hb (aHb), the posterior Hb (pHb), and the Hb enhancer classifiers independently. Approximately 40% of the sequences scored positively for at least one classifier (Supplemental Fig. S4), but only 12% (40,000) scored positively for all three (we dubbed the overlap set HbEns, as it represents the most reliable prediction of Hb enhancers). Seventy-seven of the HbEns (0.2%) are known hindbrain enhancers from the VISTA Enhancer Browser (Supplemental Table S3; Visel et al. 2007), and 26,000 (60%) overlap enhancer marks (H3K4me1 or H3K27ac, Methods). Reflecting the similarity of the training data, we observed a large overlap among the highest scoring predictions obtained by each Hb classifier (Fig. 1B). The overlap correspondingly increases with an increase in score cutoff, suggesting that sequence signatures for general Hb activity, rather than anterior or posterior Hb, dominate the decision function of all classifiers.

The genomic distribution of the HbEns is similar to that observed for the training set. Approximately half of the candidate enhancers are intronic and half are intergenic (see Supplemental custom track 1 HbEns). Also, HbEns are fairly uniformly distributed with respect to the conserved sequences that served as the basis for the genome scans, with an average of four candidates per locus and a maximum of 102 in the case of *PTPRD*, a 2.3 Mb gene highly expressed in brain and recently associated with ADHD (Elia et al. 2010). Compared with all scanned conserved sequences, HbEns are enriched within the loci of genes that are known to play a role in Hb development (P-value $= 2.8 \times 10^{-9}$, hypergeometric test) (Supplemental Table S4). Moreover, higher scoring predictions are located significantly closer to genes associated with Hb development (Fig. 1C; Supplemental Table S3), indicating that our method identifies enhancers that are active in the Hb. Although all HbEns are, by definition, conserved among mammals, their level of evolutionary conservation is notably elevated. HbEns are significantly more conserved with respect to the conserved sequences that served as the basis for the genome scans (based on average phastCons scores [Siepel et al. 2005], P-value $< 2.2 \times 10^{-16}$, Wilcoxon rank-sum test). Additionally, 21% of HbEns are shared with chicken, 8% with frog, and 3% with zebrafish (Methods). Also, with respect to the conserved sequences that served as the basis for the genome scans, conservation in vertebrates is slightly, but significantly, enriched among HbEns (P-value is 2.3×10^{-11} for the overlap with regions that are also conserved in chicken, Fisher's exact test). Moreover, we found a statistical enrichment of DNase I hypersensitive sites (HSS) identified in genomic DNA isolated from human fetal brain among HbEns (1.2-fold enrichment as compared with low-scoring sequences, P-value $< 2.2 \times 10^{-16}$, Fisher's exact test), while we do not observe any enrichment for DNase I HSS in other fetal tissues, such as heart and lung. Although, the hindbrain is only a subset of the complex tissue analyzed in fetal brain, and may refer to a different developmental stage, the enrichment in brain DNase I HSS corroborates our predictions as tissue-specific enhancers.

Finally, to evaluate the ability of our method to accurately define tissue-specific sequence patterns, we compared the distribution of predicted Hb enhancers with forebrain, midbrain, and limb enhancer predictions obtained in the same manner. In particular, we sought to verify that our predictions are not generally shared between different tissues, which would suggest a failed attempt to define a tissue-specific classifier. After we trained additional classifiers on the corresponding EP300 ChIP-seq enhancer sets, we found that there is <20% overlap between the top 5% of high scoring predictions (16% forebrain, 13% midbrain, 9% limb). This overlap is further reduced to 12% when comparing the top 1% of high scoring predictions. This confirms our hypothesis that genome-wide predictions of classifiers trained on enhancers with different activities constitute largely disjoint sets, suggesting that the corresponding classifiers recognize sequence patterns linked to different biological functions.

The hindbrain classifier is a highly accurate predictor of hindbrain activity in zebrafish

In order to determine the accuracy of our method we set out to determine how frequently our predictions identify active Hb enhancers in vivo. In total we selected 55 sequences with a positive scaled summary Hb (see Methods) for functional evaluation in a zebrafish transgenic reporter assay (Supplemental Table S5). To avoid the introduction of biases based on genomic position, we included both intronic and intergenic sequences residing on 21 different human chromosomes (all except chr10 and Y) (Supplemental Table S5). In addition, six low scoring sequences with a scaled summary Hb score less than zero were selected as likely "negative" predictions. Predicted sequences may not identify all functional components within a complete enhancer, thus although our predictions were based on 100–200 bp sequence intervals, we designed primers to include ~200 bp flanking each side of the original sequence. The average size of all assayed amplicons was 485 bp (Supplemental Table S6).

All sequences were tested for enhancer activity in the Hb using our established zebrafish transgenesis pipeline (Fisher et al. 2006; McGaughey et al. 2008). We define hindbrain expression as any expression in the CNS region that is posterior to the midbrain

extending through the myelencephalon and delimited by the anterior portion of the spinal cord. Since the training set sequences directed expression in a number of non-Hb tissues, we do not require that expression is restricted to this region, and are therefore testing the sensitivity of the classifiers to Hb patterns rather than the specificity. The vast majority of constructs (51/55 putative Hb enhancers) directed reporter expression in some portion of the CNS in mosaic zebrafish embryos. Similarly, 6/6 low likelihood predicted sequences displayed mosaic signals in some part of the embryo, including portions overlapping the CNS. All embryos that displayed reporter expression in mosaics were raised to maturity and crossed with AB zebrafish to determine which sequences could direct EGFP expression in the Hb. In total we identified two or more founders for 34 putative Hb enhancers, of which 30 (88%) founder sets displayed concordant expression in the Hb (Supplemental Fig. S5), a predictive success rate that exceeds prior rates of enhancer validation using both computational predictions as well as EP300 ChIP-seq-based predictions ([Narlikar et al. 2010], 62% [Blow et al. 2010], 84%, Fig. 2). In contrast, none of the six low likelihood controls displayed consistent expression in the Hb when passed though the germline (Supplemental Fig. S5).

The patterns observed in stable lines displayed marked pleiotropy in their range of reporter expression both in Hb regions as well as in non-Hb regions, likely reflecting the heterogeneity of the training set. Figure 3 provides eight examples that illustrate the diverse patterns of expression observed in our validation set. Although the models trained on anterior and posterior sets of sequences did not appear to be particularly predictive of the relative position of Hb expression, we found that the resulting patterns could be grouped into categories displaying similar expression. HB41, HB34, and HB02 share an expression profile that includes the cerebellum, part of the anterior Hb, in addition to varying levels

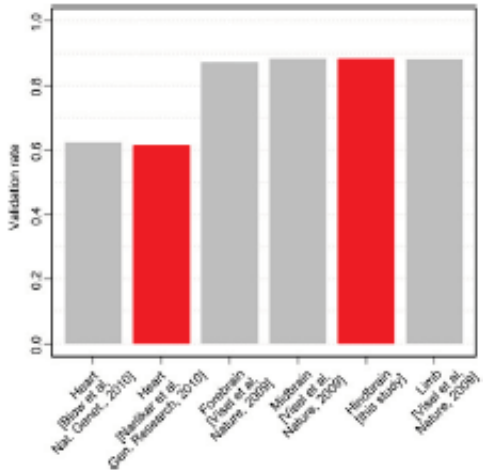


Figure 2. Experimental validation of tissue-specific enhancer candidates in transgenic zebrafish and mouse assays. Our computational approach trained on small empirical data sets (red bars) resulted in validation rates comparable to those for ChIP-seq-derived data sets using an EP300 antibody (gray bars) for the heart. Similarly, the validation rates of computational Hb classifiers trained on small empirical data sets were also comparable to those obtained with EP300 ChIP-seq experiments in other brain tissues.

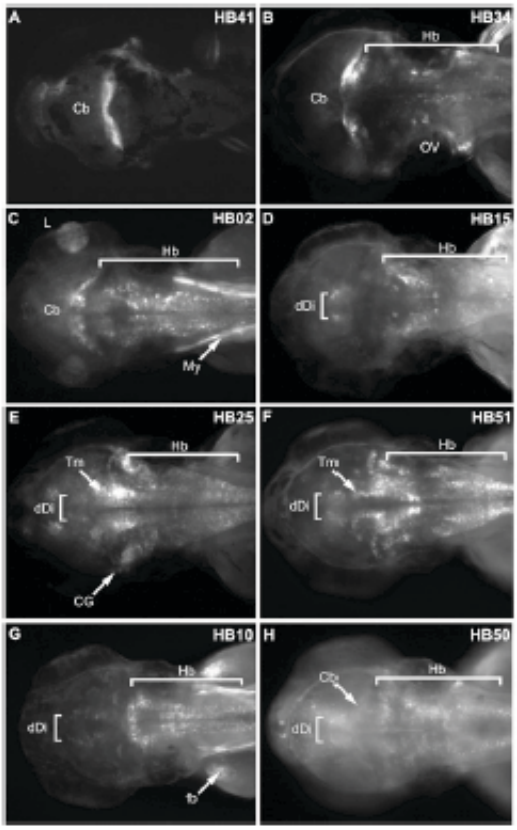


Figure 3. Predicted enhancers display pleiotropic expression patterns in the hindbrain. (*A–H*) GFP reporter expression from eight stable lines corresponding to Hb predictions showing expression across the Hb as well as in some non-Hb domains. Dorsal view images were taken at 3 dpf (for lateral images, see Supplemental Figures), anterior to the *left*. (*A*) HB41, (*B*) HB34, (*C*) HB02, (*D*) HB15, (*E*) HB25, (*F*) HB51, (*G*) Hb10, (*H*) HB50. (Cb) cerebellum; (OV) otic vesicle; (Hb) hindbrain; (L) lens; (My) myotome; (dDi) dorsal diencephalon; (Tm) tegmentum; (CG) cranial ganglia; (fb) fin bud.

of expression along the length of the Hb (Fig. 3A–C). However, HB02 also directs non-neuronal expression in the lens of the eye and myotome (Fig. 3C), which may be a result of position effects based on the site of amplicon insertion in the zebrafish genome as it was not observed in all stable lines. Some sequences, like HB15 (Fig. 3D), show expression in the Hb and very little extraneous expression. In contrast, HB25 and 51 share a different expression profile displaying strong expression in the dorsal Hb as well as the tegmentum, a structure in the midbrain that is continuous with the medulla oblongata (Fig. 3E,F; Thisse and Thisse 2004; Thisse et al. 2004). HB10 shows distinct expression in the Hb, spinal cord, and dorsal diencephalon, as well as faint expression in the tegmentum and non-neuronal expression in the myotome and fin buds (Fig. 3G). In contrast to the distinct Hb expression seen in HB10, many domains within the CNS are faintly marked by reporter expression directed by HB50, including Hb neurons, cerebellum,

tegmentum, dorsal diencephalon, and telencephalon (Fig. 3H). The varied patterns of expression observed within the Hb validation set are consistent with the diverse nature of the motifs comprising the classifier. This is expected given that the training set is comprised of sequences that displayed significant pleiotropy and included sequences that directed expression in an array of Hb subdomains, as well as in non-Hb tissues. Consequently, we expected that TFBSs contained within these amplicons, and contributing to their prediction, would be diverse. However, we also anticipated that they would include sites for factors whose endogenous expression overlap with domains of reporter expression.

In vivo validated Hb enhancer sequences are enriched for the 100 most relevant motifs for discriminating Hb enhancers compared with random sequences with similar GC content (Supplemental Table S7). TFBSs for proteins in the POU, NKX, or PAX families, as well as LHX3 are especially common in our validation set (Supplemental Table S8). Consistent with the in silico evaluations of TFBSs identified in HbEns collectively, factors in these families play critical roles in neuronal development. Furthermore, the observed reporter expression for each is largely consistent with previously published expression patterns for one or more of the corresponding TFs. POU domains are found in a large family of TFs and bind the consensus sequence ATGCAAAT (Verrijzer and Van der Vliet 1993). They are expressed mainly in the CNS, and act as regulators of neurogenesis in zebrafish (Spaniol et al. 1996). Consistent with these data, POU family TFBSs were the most commonly identified sites in our validation set and showed an enrichment of 2.6 over GC matched control sequences (P-value = 0.01, Fisher's exact test) (Supplemental Table S8) and many of our elements share expression domains with POU factors. NKX proteins are necessary for the proper development of motor neurons in the hindbrain (Pattyn et al. 2003) and consistent with this role we see a significant enrichment (2.4, P-value = 0.005, Fisher's exact test) for NKX family TFBSs in our validated set of Hb enhancers. Similarly, the PAX gene family similarly comprises a large group of highly conserved TFs required for neuronal development (Wang et al. 2010; Thompson and Ziman 2011). Furthermore, 10/30 validated predictions contained at least one PAX family motif (enrichment of 1.8 as compared with random genomic sequences with similar length and GC-content, P-value = 0.05, Fisher's exact test). Finally, a number of sequences share in common an LHX3 motif that binds a LIM domain TF with a role in neuronal specification (Cepeda-Nieto et al. 2005; Gadd et al. 2011), resulting in an enrichment of 4.6 (P-value = 0.0002, Fisher's exact test). HB25 and HB51 both contain an LHX3 TFBS and share many overlapping domains of reporter expression, including in the Hb and spinal column, which is consistent with endogenous lhx3 expression (Fig. 3E,F; Supplemental Fig. S5; Thisse and Thisse 2004; Thisse et al. 2004). In contrast to the HbEns sequences tested, only one of the low likelihood controls contained any of these motifs, supporting their high predictive power in our model. Taken collectively, these data provide compelling evidence that the validated sequences may play important roles in regulating transcription in the developing Hb.

Enhancer activity is due to the presence of specific transcription factor motifs

Our data suggest that TFBSs contributing to the classifier might independently or collectively explain aspects of the observed regulatory control of the sequences within which they reside. We selected two sequences with Hb regulatory control (HB01 and HB16) to examine more closely, surveying the distribution of TFBSs within each predicted sequence. We then identified smaller sequence fragments for analysis in zebrafish based on the clustering of TFBSs therein.

The full-length HB01 sequence directed distinct expression in the rhombomeres, as well as the midbrain Hb boundary, cranial ganglia, and dorsal diencephalon (Fig. 4B,E). We amplified two smaller fragments (HB01_I and HB01_II) from within this full-length sequence based on the pattern of TFBSs clusters. HB01_I is a 56-bp sequence containing motifs for PITX2, CDX, CEBPG, NKX3-1, and BCL6 (Fig. 4A). Upon passage through the germline, HB01_I displayed broad reporter expression in the CNS (Fig. 4C,F). This pattern encompassed the expression domains marked by the full-length HB01. The expanded expression domains marked by HB01_I could reflect the increased efficiency of TFBSs being placed closer to the minimal promoter (Nolis et al. 2009). It may also reflect the absence of other regulatory sequence motifs within or beyond the initial predicted interval which otherwise act in the full-length construct to moderate transcriptional activity (Gompel et al. 2005). Notably HB01_II, which is 93 bp and contains motifs for HNF3, POU2F1, NKX2-5, MYOG, SOX10, and HMGA1 (Fig. 4A), did not show any mosaic expression, and was determined to be insufficient for enhancer activity in the Hb in this assay (Fig. 4D,G).

Similarly, HB16 displays prominent expression in the dorsal Hb and faint expression in the ventral Hb and lateral tegmentum (Fig. 4I,M). Once again we amplified three short fragments from within the initially predicted sequence based on TFBS clusters (Fig. 4H). HB16_I is a 29-bp fragment containing a GATA1 motif; HB16_II is 54 bp in length and contains MEOX2, NKX6-1, EN1, TAL1, NKX2 family, JUN, and PAX4 motifs; and HB16_III is a 23-bp fragment encompassing a HOXA4 motif (Fig. 4H). Upon passage through the germline, both HB16_I and HB16_II directed expression in the Hb (Fig. 4J,K,N,O). In contrast, HB16_III only drove expression in the myotome of stable lines (Fig. 4L,P).

Notably, the reporter expression in the Hb neurons and the lateral tegmentum directed by HB16_I are similar to those of the endogenous *gata3* (Fig. 4J,N; Thisse and Thisse 2004; Thisse et al. 2004). This pattern is further consistent with expression directed by full-length HB16. Furthermore, HB16_II directs expression along the entire length of the ventral and medial Hb and spinal column (Fig. 4K,O). As such, it overlaps much of the Hb domain marked by HB16 and resembles the endogenous expression of *nkx6* family, *nkx2* family, and *tal1* RNA (Thisse and Thisse 2004; Thisse et al. 2004; Binot et al. 2010). The observed tegmental reporter expression is also consistent with endogenous expression of *tal1* (Thisse and Thisse 2004; Thisse et al. 2004). A potential role for the JUN TFBS identified in HB16 is not immediately obvious but these factors display much broader expression domains throughout the CNS and may in part account for expression domains extending dorsally. Although not conclusive, these data suggest that the expression of TFs corresponding to motifs contributing to our classifier are consistent with their predicted biological roles in modulating expression in the Hb and show that enhancers can be further broken down into their component TFBS fragments and continue to faithfully drive reporter expression in the predicted tissue.

Discussion

The exquisite orchestration of transcriptional control is essential for the normal development and homeostasis of multicellular organisms. Systematic identification of sequences responsible for these activities, however, has proven a significant challenge. Although

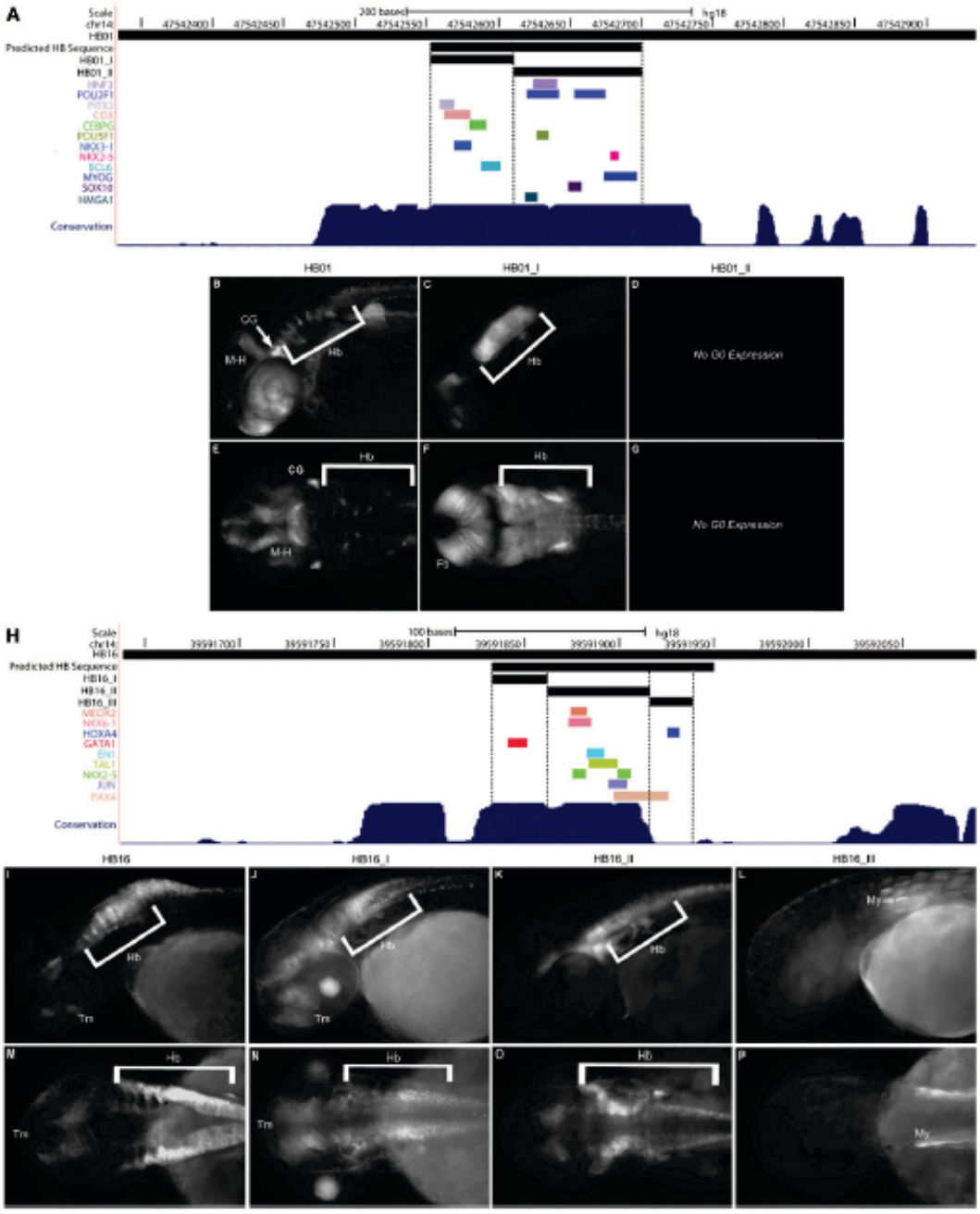


Figure 4. TF clustering reveals functional sequence domains. (*A*,*H*) UCSC Genome Browser custom track showing injected construct, classifier predicted HB sequence, and fragments tested for Hb expression (black bars, *top* to *bottom*). Colored bars mark TFBS for various factors. (*B*–*G*, *I*–*P*) GFP reporter expression observed with each sequence (lateral view, *top*; dorsal view, *bottom*). All images taken at 2 dpf, anterior to the *left*. (*A*) HB01 custom track with two subcloned fragments, (*B*) full-length HB01, lateral view, (*C*) HB01_I, lateral view, (*D*) HB01_II, no G0 GFP reporter expression observed, (*E*) full-length HB01, dorsal view, (*F*) HB01_I, dorsal view, (*G*) HB01_II, no G0 GFP reporter expression observed. (*H*) HB16 custom track with three subcloned fragments, (*I*) full-length HB16, lateral view, (*J*) HB16_I, lateral view, (*K*) HB16_II, lateral view, (*L*) HB16_III, lateral view, (*M*) full-length HB16, dorsal view, (*N*) HB16_I, dorsal view, (*O*) HB16_II, dorsal view, (*P*) HB16_III, dorsal view. (CG) cranial ganglia; (M-H) midbrain hindbrain boundary; (Hb) hindbrain; (Fb) forebrain; (Tm) tegmentum; (My) myotome; (L) lens.

the encryption of regulatory instructions in DNA sequence is well established, the absence of an established vocabulary has precluded the prediction of biological activities rendered by noncoding functional genome components based on inspection of the primary sequence. Two strategies commonly employed in the identification of transcriptional enhancers are evolutionary sequence constraint and ChIP-seq. Although sequence constraint has been used with significant success, it can impute little regarding the likely biological activity of any identified sequence. Similarly ChIP-seq profiling of TFs, histone modifications, and transcriptional co-activators such as EP300 has recently emerged as a powerful tool for the identification of enhancers active in various tissues; however, not all enhancers are captured by affinity-based methods, and not all cell types are amenable to these assays. Recent efforts to identify sequence motifs (active TFBS) have proven increasingly powerful, allowing the elucidation of early language structure for regulatory control in specific tissues (Narlikar et al. 2010; Lee et al. 2011).

We have integrated these computational strategies, employing machine learning to train a sequence-based classifier on a set of largely published in vivo validated enhancers in the Hb. The result is a highly accurate predictor of enhancer activity in the Hb. When applied to the human genome, 88% (30/34) of sequences demonstrate Hb regulatory control when assayed in vivo (stable zebrafish transgenesis). In contrast, even among sequences identified as being deeply conserved only ~8% were observed to drive expression in the Hb (Pennacchio et al. 2006). The motifs identified by our classifier frequently represent TFBSs for factors with known roles in regulating transcription in the Hb and with endogenous expression patterns overlapping with that of reporter expression. Furthermore, we show that, consistent with our classifier, clusters of TFBS (~30–100 bp) contributing to predicted Hb regulatory control can account for aspects of Hb regulatory expression observed in the original (~500 bp) sequence from which they were derived.

Although the vocabulary described is an effective predictor of Hb activity, we observed pleiotropy among Hb domains marked by reporter expression as well as expression in domains outside the Hb, including non-neural tissues. These observations are consistent with the complexity of vertebrate enhancers known to display a broad expression pattern across multiple tissues (Visel et al. 2007). It is particularly important to keep in mind that the Hb enhancers in our training data set were not exclusively expressed in the Hb, but largely displayed multi-tissue expression patterns. From the sequence analysis perspective, our training set contained a large group of Hb enhancers and several smaller clusters of other expression subdomains. All non-Hb signatures in our training created a plethora of misleading signals confusing the classifier. However, the high Hb validation rate of HbEns reflects the ability of the classifier to sensitively extract the Hb sequence encryption from the noisy input data set. Knowing that Hb sequence encryption often resides within enhancers with broad expression patterns and does not represent a code of exclusive Hb expression, we were not surprised to observe that expression is not specific to Hb in experimentally validated HbEns.

As additional support for the utility of our model, we find that our predicted Hb enhancers are enriched for a particularly large number of CNS TFBSs compared with TFs known to be active in other tissues. Our experimental data also suggest that Hb enhancers can be divided into independent functional subunits, here tested as TFBS clusters, with similar activities but different sequence structures—an observation that highlights flexibility of the Hb sequence encryption with potential for adaptation to additional functions and the use of different activation mechanisms. The observed biological behaviors of these TFBS clusters were consistent with the known patterns of expression of TF family members predicted to bind them. This raises the possibility that retraining algorithms using subsets of training or predicted sequence sets may define the sequence grammar specific to individual Hb sub-domains and cell types.

Computational methods are becoming increasingly powerful tools for enhancer prediction. Experimental validation rates for computer learning algorithms are comparable to those achieved by experimental ChIP-seq predictions and can be similarly independently correlated with the presence of features known to be present in active enhancers such as known TFBS motifs, specific histone marks, and increased conservation. This study demonstrates that, in addition to the sequence substrate provided by genome-wide ChIP-based strategies, the published literature may serve as a valuable entry point for such analyses of regulatory elements. We demonstrate that even a relatively small curated experimental data set can provide significant insight into the regulatory lexicon of a highly complex anatomical structure like the Hb, and that this vocabulary can likely be dissected and improved in subsequent cycles of investigation and/or by the refinement of the substrate on which it is trained. Therefore, this study adds to the ongoing project of genome annotation by identifying sequences that have a functional role in the Hb. The development of regulatory language is a pivotal step in the prediction of functional variation by inspection of the primary sequence and as such this study makes a significant first step in the development of a Hb lexicon.

Methods

Tissue-specific enhancer models

We extracted 771 human sequences from the VISTA Enhancer Browser (Visel et al. 2007) with validated in vivo enhancer activity in 23 tissues. We were able to retrieve at least 29 sequences each for 11 of these tissues.

Hindbrain enhancer models for mouse, chicken, frog, and zebrafish

Orthologous regions of the human Hb enhancer training set were identified using the liftOver utility from the UCSC Genome Browser (Karolchik et al. 2008). We discarded mapped sequences longer than 5 kb. We successfully mapped 100%, 86%, 74%, and 47% of the 211 Hb enhancers onto the mouse (mm9), chicken (galGal3), frog (xenTro2), and zebrafish (danRer5) genomes respectively (See Supplemental custom tracks 2–5).

Forebrain, midbrain, and limb enhancers identified using ChIP-seq

Genomic regions enriched for EP300 binding in mouse forebrain, midbrain, and limb tissues were extracted from Supplemental Tables 2–4 of Blow et al. (2010). We identified orthologous regions of the mouse coordinates with the liftOver utility from the UCSC Genome Browser (Karolchik et al. 2008). Sequences longer than 1 kb were discarded, resulting in a total of 2199 forebrain sequences, 1909 midbrain sequences, and 3155 limb sequences.

Background genomic sequences

For each enhancer in the training set, 10 controls with similar length, GC, and repeat-content were randomly drawn from the noncoding portion of the corresponding genome.

TFBS mapping

Putative TFBSs were identified by searching the sequences with MAST (Bailey and Elkan 1994) for 775 motifs in TRANSFAC Release 2009.2 (Matys et al. 2006) and JASPAR (Bryne et al. 2008). MAST was run independently on each individual sequence with default setup and parameters.

TF binding to de novo motifs

The identity of the TFs binding to the de novo motifs was queried using STAMP (Mahony and Benos 2007) and JASPAR (Bryne et al. 2008).

Association between TFs and TFBSs

TF annotation for known TFBSs was obtained from TRANSFAC, JASPAR, and the Broad Institute MsigDB database (Subramanian et al. 2005).

TFBS enrichment

Overrepresented TFBSs were determined by comparing the occurrence of the motifs among query sequences and background genomic sequence, and applying Fisher's exact test. We used a *P*-value threshold of 0.05. When indicated, we adjusted the *P*-values for multiple testing using the procedure suggested by Benjamini and Hochberg (1995).

Enhancer models

Each enhancer model was trained to distinguish between enhancers specific for a given tissue and other noncoding sequences, randomly drawn from the noncoding human sequence, with length, GC, and repeat-content distributions similar as those observed for the enhancers. The decision of the corresponding classifier was based on the presence or absence of two different types of motifs: 775 corresponding to binding specificities of vertebrate TFs compiled in public databases (TRANSFAC and JASPAR [Matys et al. 2003; Bryne et al. 2008]), and 20 short sequence patterns enriched among the set of enhancers, identified with PRIORITY (Narlikar et al. 2007), which should account for the binding of unknown TFs or TFs with unknown binding specificities. Thus, each sequence was represented as a feature vector indicating the number of matches per base pair to each of these motifs, computed using MAST (Bailey and Gribskov 1998). We built the classifier using linear SVMs (implemented in libsvm [Chang and Lin 2011]), assuming no prior knowledge of TFs active in the different tissues, with the goal being to discover them using the feature weights learned by the classifier.

Extracting homogeneous Hb enhancer data sets

Hb enhancers tend to drive expression in multiple tissues, and even show heterogeneous patterns of expression within the Hb. As a result it is unlikely that we would be able to identify a unique set of sequence features representing all Hb enhancers. Thus, similarly to the approach taken in Narlikar et al. (2010), we selected a large subset of these sequences sharing homogeneous sequence features as an attempt to reduce the sequence heterogeneity among the 212 human Hb enhancers. For this purpose, we repeated the 10-fold cross-validation on five random partitions of the Hb enhancer data set as well as on that of the corresponding controls, and selected only those Hb enhancers that were classified as such in at least 50% of the times in which they were tested for the final training set. Therefore, the final human Hb enhancer data set contained 124 sequences.

Performance assessment of enhancer classifiers

The performance of the classifiers was evaluated in a 10-fold cross-validation, using the area under the ROC curve (AUC). AUC values range from 0.5 (random discrimination) to a theoretical maximum of 1.

Linear SVMs

Training a linear SVM classifier is equivalent to solving the following constrained optimization problem (Shawe-Taylor and Cristianini 2002):

Given the training samples $T = \{(x_i, y_i) | x_i \in \mathbb{R}^p, y_i \in \{-1,1\}\}_{i=1}^n$, find the values of w, b and ξ_i that minimize

$$\frac{1}{2} w^T \cdot w + C \sum_{i=1}^{n} \xi_i$$

satisfying the constraints

$$w_j \geq 0$$

and

$$\xi_i \geq 0 \; \forall i = 1, \ldots, n.$$

The decision function of the classifier for an unknown sample x is given by

$$w_j < 0.$$

The dual form of this problem is:

Given the training samples $T = \{(x_i, y_i) | x_i \in \mathbb{R}^p, y_i \in \{-1,1\}\}_{i=1}^n$, find the values $\{\alpha_i\}_{i=1}^n$ that maximize

$$\sum_i \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

satisfying the constraints

$$0 \leq \alpha_i \leq C \; \forall i = 1, \ldots, n$$

and

$$\sum_{i=1}^{n} a_i y_i = 0.$$

Samples x_i for which $a_i \geq 0$ are called support vectors.

The vector w can be computed in terms of α_i as

$$w = \sum_{i=1}^{n} a_i y_i x_i$$

and, therefore, contains the weighted features of the support vectors.

SVM parameter selection

Linear SVMs have only one parameter, C, which controls the trade-off between errors on the training data and margin maximization. We found that the performance of the Hb enhancer classifier was relatively stable with respect to changes in C. We estimated C based on the training data as $\left[\frac{1}{n}\sum_{i=1}^{n} |x_i|\right]^{-2}$.

Additionally, because the training data are unbalanced (there are 10 controls for each enhancer sequence), misclassifications are penalized differently depending on the class of sequences (controls and enhancers), proportionally to the total number of sequences in each class.

Motif rankings

After obtaining a linear SVM model, the weight vector *w* can be used to decide the relevance of each feature (Guyon et al. 2002).

The larger $|w_j|$, the more important role of feature j in the decision function. We rank features—in our case, motifs—according to $|w_j|$. We exclude de novo motifs from these ranks unless stated otherwise. It is important to note that this interpretation for w is only valid for linear SVMs.

Hindbrain genes

We identified a set of 787 human genes likely to be involved in Hb function by retrieving genes with relevant phenotypes from the Online Mendelian Inheritance in Man (OMIM) database (Amberger et al. 2009) and the corresponding orthologs of genes with pertinent annotation in the Mammalian Phenotype (MP) Browser at the Mouse Genome Informatics website, The Jackson Laboratory, Bar Harbor, Maine (http://www.informatics.jax.org).

Genome scans

We applied three human Hb enhancer models trained on (1) the complete Hb data set, (2) the subset of Hb enhancers that are active in the anterior Hb, and (3) the subset of enhancers which functions in the posterior Hb to scan sequences highly conserved across mammals using the Most Conserved Elements database from the UCSC Table Browser (Siepel et al. 2005). Noncoding conserved sequences were determined based on annotation in UCSC Known and RefSeq (Hsu et al. 2006; Pruitt et al. 2009). Sequences within 100 bp of each other were clustered together and clusters <100 bp were excluded from the analysis. Using classifiers trained on the orthologous sequences of the complete data set of human Hb enhancers, we utilized an analogous procedure to predict Hb-specific enhancers in the mouse, chicken, frog, and zebrafish genomes.

Scaled summary Hb score

Each scanned sequence is given three scores, $score_{anterior\_Hb}$, $score_{posterior\_Hb}$, and $score_{general\_Hb}$, by the classifiers trained on the subset of Hb enhancers that are active in the anterior Hb, the subset of enhancers which functions in the posterior Hb, and the complete Hb data set, respectively. The scores are distributed in the range [−17,15], [−20,15], and [−22,15], respectively (see Supplemental Fig. S4). Scores >0 correspond to putative enhancers active in the anterior Hb, in the posterior Hb, and in the (general) Hb, respectively. Approximately 130,000 sequences scored >0 for at least one of the classifiers, while 40,000 sequences scored >0 for all three. Scores for all classifiers are subsequently linearly scaled according to

$$score^* = \begin{cases} -\left(1 - \frac{score - score_{min}}{-score_{min}}\right), & if\ score < 0 \\ \frac{score}{1 + score_{max}}, & if\ score \geq 0 \end{cases},$$

where $score_{min}$ and $score_{max}$ are the minimum and maximum scores obtained in the genome-wide scan, respectively.

Finally, we define the scaled summary Hb score as the maximum between $score^*_{anterior\_Hb}$, $score^*_{posterior\_Hb}$, and $score^*_{general\_Hb}$.

Association between enhancer predictions and loci

For defining gene loci in the human genome, we used the knownGene and RefSeq annotation tracks available at the UCSC Genome Browser (November 2011). Each locus was defined by one or more overlapping transcripts, prohibiting overlap among different loci. Putative Hb enhancers were associated with loci based on genomic proximity. Thus, each putative Hb enhancer is assumed to target the genes in the nearest locus.

DNase I hypersensitivity

We compared our putative Hb enhancers with human fetal brain, heart, and lung DNase I hypersensitive peaks from http://nihroadmap.nih.gov/epigenomics/.

H3K4me1 and H3K27ac

H3K4me1 and H3K27ac peaks were downloaded from http://genome.ucsc.edu/ENCODE/ (The ENCODE Project Consortium 2011) and correspond to multiple human cell lines (all available to date).

In vivo validation

Candidate Hb enhancers for validation were selected randomly from positively scoring sequences with rank less than or equal to ~40,000. Controls were selected among sequences that scored among the bottom 1% (i.e., rank greater than or equal to ~334,000) for all classifiers. Zebrafish were maintained as previously described (Kimmel et al. 1995; Westerfield 2000). Predicted enhancers were amplified by PCR from human genomic DNA and cloned using Gateway Technology (Invitrogen). PCR fragments were TA-cloned into the pCR8/GW/TOPO vector (Invitrogen) then TOPO-cloned using attL1 and attL2 sites into the pT2cfosGW vector for injection into zebrafish embryos. Short fragment sequences for HB01 and HB16 were synthesized as double-stranded oligos, A overhangs added, then cloned as predicted enhancers. At least 100 embryos were injected per construct at the two-cell stage with tol2 transposase as previously described (Fisher et al. 2006). Injected embryos were screened for GFP expression in the CNS at 24 and 48 hpf. Those showing CNS expression were raised to adulthood and crossed to AB zebrafish. G1 embryos were screened for Hb expression at 24, 48, and 72 hpf. GFP positive embryos were live-imaged at 72 hpf using a Carl Zeiss Lumar V12 Stereo microscope with AxioVision version 4.8 software. Embryos were fixed in 4% PFA (Sigma) overnight then post-fixed in 100% acetone (JT Baker) and washed in PBS with 0.5% Tween. Embryos were blocked in 10% goat serum and 1% BSA for two hours, then incubated with chicken anti-GFP (Invitrogen A10262, 1:1000) overnight. After washing, Alexa Fluor 488 goat anti-chicken IgG (Invitrogen A11039, 1:3000) was added and incubated overnight. After washing, embryos were stored in 80% glycerol at 4°C for future imaging.

Acknowledgments

This work was funded in part by the National Institute of Neurological Disease and Stroke (NS062972) to A.S.M., the Intramural Research Program of the NIH, National Library of Medicine to I.O., and a predoctoral training grant (GM07814) to X.R.

Author contributions: This study was conceived by A.S.M. and I.O. The zebrafish transgenic analyses were designed and executed by G.M.B., X.R., T.M., and A.S.M. The development of the Hb classifier and computational analyses were performed by L.T. and I.O. Z.E.S. computed the enrichment in DNaseI HS. All authors contributed to the interpretation of experimental data. The manuscript was written by A.S.M., L.T., I.O., X.R., G.M.B., and Z.E.S.

References

Amberger J, Bocchini CA, Scott AF, Hamosh A. 2009. McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res* **37**: D793–D796.

Andreasen NC, Pierson R. 2008. The role of the cerebellum in schizophrenia. *Biol Psychiatry* **64**: 81–88.

Aston-Jones G. 2005. Brain structures and receptors involved in alertness. *Sleep Med* (Suppl 1) **6:** S3–S7.
Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2:** 28–36.
Bailey TL, Gribskov M. 1998. Methods and statistics for combining motif match scores. *J Comput Biol* **5:** 211–221.
Banerji J, Rusconi S, Schaffner W. 1981. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27:** 299–308.
Bardet PL, Schubert M, Horard B, Holland LZ, Laudet V, Holland ND, Vanacker JM. 2005. Expression of estrogen-receptor related receptors in amphioxus and zebrafish: Implications for the evolution of posterior brain segmentation at the invertebrate-to-vertebrate transition. *Evol Dev* **7:** 223–233.
Barski A, Zhao K. 2009. Genomic location analysis by ChIP-Seq. *J Cell Biochem* **107:** 11–18.
Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol* **57:** 289–300.
Bequin PC, Gledd JN, Jacobsen LK, Hamburger SD, Krain AL, Rapoport JL, Castellanos FX. 1998. Cerebellum in attention-deficit hyperactivity disorder: A morphometric MRI study. *Neurology* **50:** 1087–1093.
Binot AC, Manfroid I, Flasse L, Winandy M, Motte P, Martial JA, Peers B, Voz ML. 2010. Nkx6.1 and nkx6.2 regulate α- and β-cell formation in zebrafish by acting on pancreatic endocrine progenitor cells. *Dev Biol* **340:** 397–407.
Blow MJ, McCulley DJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, et al. 2010. ChIP-Seq identification of weakly conserved heart enhancers. *Nat Genet* **42:** 806–810.
Bryne JC, Valen E, Tang MH, Marstrand T, Winther O, da Piedade I, Krogh A, Lenhard B, Sandelin A. 2008. JASPAR, the open access database of transcription factor-binding profiles: New content and tools in the 2008 update. *Nucleic Acids Res* **36:** D102–D106.
Bulger M, Groudine M. 2011. Functional and mechanistic diversity of distal transcription enhancers. *Cell* **144:** 327–339.
Cepeda-Nieto AC, Pfaff SL, Varela-Echavarria A. 2005. Homeodomain transcription factors in the development of subsets of hindbrain reticulospinal neurons. *Mol Cell Neurosci* **28:** 30–41.
Chang C-C, Lin C-J. 2011. LIBSVM: A library for support vector machines. *ACM Trans Intell Syst Technol* **2:** article 27. doi: 10.1145/1961189.1961199.
Cooper GM, Shendure J. 2011. Needles in stacks of needles: Finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet* **12:** 628–640.
Elia J, Gai X, Xie HM, Perin JC, Geiger E, Glessner JT, D'Arcy M, deBerardinis R, Frackelton E, Kim C, et al. 2010. Rare structural variants found in attention-deficit hyperactivity disorder are preferentially associated with neurodevelopmental genes. *Mol Psychiatry* **15:** 637–646.
The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447:** 799–816.
The ENCODE Project Consortium. 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* **9:** e1001046. doi: 10.1371/journal.pbio.1001046.
Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, et al. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473:** 43–49.
Fisher S, Grice EA, Vinton RM, Bessling SL, McCallion AS. 2006. Conservation of RET regulatory function from human to zebrafish without sequence similarity. *Science* **312:** 276–279.
Gadd MS, Bhati M, Jeffries CM, Langley DB, Trewhella J, Guss JM, Matthews JM. 2011. Structural basis for partial redundancy in a class of transcription factors, the LIM homeodomain proteins, in neural cell type specification. *J Biol Chem* **286:** 42971–42980.
Ghysen A. 2003. The origin and evolution of the nervous system. *Int J Dev Biol* **47:** 555–562.
Giguere V. 2002. To ERR in the estrogen pathway. *Trends Endocrinol Metab* **13:** 220–225.
Gompel N, Prud'homme B, Wittkopp PJ, Kassner VA, Carroll SB. 2005. Chance caught on the wing: *cis*-regulatory evolution and the origin of pigment patterns in *Drosophila*. *Nature* **433:** 481–487.
Guyon I, Weston J, Barnhill S, Vapnik V. 2002. Gene selection for cancer classification using support vector machines. *Mach Learn* **46:** 389–422.
He A, Kong SW, Ma Q, Pu WT. 2011. Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart. *Proc Natl Acad Sci* **108:** 5632–5637.
Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D. 2006. The UCSC known genes. *Bioinformatics* **22:** 1036–1046.
Karolchik D, Kuhn RM, Baertsch R, Barber GP, Clawson H, Diekhans M, Giardine B, Harte RA, Hinrichs AS, Hsu F, et al. 2008. The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res* **36:** D773–D779.
Kelberman D, de Castro SC, Huang S, Crolla JA, Palmer R, Gregory JW, Taylor D, Cavallo L, Faienza MF, Fischetto R, et al. 2008. SOX2 plays a critical role in the pituitary, forebrain, and eye during human embryonic development. *J Clin Endocrinol Metab* **93:** 1865–1873.
Kimmel CB, Ballard WW, Kimmel SR, Ullmann B, Schilling TF. 1995. Stages of embryonic development of the zebrafish. *Dev Dyn* **203:** 253–310.
Kittler J, Hatef M, Duin RPW, Matas J. 1998. On combining classifiers. *IEEE Trans Pattern Anal Mach Intell* **20:** 226–239.
Klyota T, Kato A, Altmann CR, Kato Y. 2008. The POU homeobox protein Oct-1 regulates radial glia formation downstream of Notch signaling. *Dev Biol* **315:** 579–592.
Lee D, Karchin R, Beer MA. 2011. Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Res* **21:** 2167–2180.
Lettice LA, Heaney SJ, Purdie LA, Li L, de Beer P, Oostra BA, Goode D, Elgar G, Hill RE, de Graaff E. 2003. A long-range *Shh* enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet* **12:** 1725–1735.
Mahony S, Benos PV. 2007. STAMP: A web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res* **35:** W253–W258.
Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, et al. 2003. TRANSFAC: Transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* **31:** 374–378.
Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, et al. 2006. TRANSFAC and its module TRANSCompel: Transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* **34:** D108–D110.
McGaughey DM, Vinton RM, Huynh J, Al-Saif A, Beer MA, McCallion AS. 2008. Metrics of sequence constraint overlook regulatory sequences in an exhaustive analysis at *phox2b*. *Genome Res* **18:** 252–260.
Narlikar L, Gordan R, Hartemink AJ. 2007. A nucleosome-guided map of transcription factor binding sites in yeast. *PLoS Comput Biol* **3:** e215. doi: 10.1371/journal.pcbi.0030215.
Narlikar L, Sakabe NJ, Blanski AA, Arimura FE, Westlund JM, Nobrega MA, Ovcharenko I. 2010. Genome-wide discovery of human heart enhancers. *Genome Res* **20:** 381–392.
Nelson SB, Janiesch C, Sander M. 2005. Expression of *Nkx6* genes in the hindbrain and gut of the developing mouse. *J Histochem Cytochem* **53:** 787–790.
Nolis IK, McKay DJ, Mantouvalou E, Lomvardas S, Merika M, Thanos D. 2009. Transcription factors mediate long-range enhancer-promoter interactions. *Proc Natl Acad Sci* **106:** 20222–20227.
Noonan JP, McCallion AS. 2010. Genomics of long-range regulatory elements. *Annu Rev Genomics Hum Genet* **11:** 1–23.
Pattyn A, Vallstedt A, Dias JM, Sander M, Ericson J. 2003. Complementary roles for Nkx6 and Nkx2 class proteins in the establishment of motoneuron identity in the hindbrain. *Development* **130:** 4149–4159.
Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, et al. 2006. In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444:** 499–502.
Pruitt KD, Tatusova T, Klimke W, Maglott DR. 2009. NCBI Reference Sequences: Current status, policy and new initiatives. *Nucleic Acids Res* **37:** D32–D36.
Schneider-Maunoury S, Gilardi-Hebenstreit P, Charnay P. 1998. How to build a vertebrate hindbrain. Lessons from genetics. *C R Acad Sci III* **321:** 819–834.
Shawe-Taylor J, Cristianini N. 2002. On the generalisation of soft margin algorithms. *IEEE Trans Inf Theory* **48:** 2721–2735.
Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15:** 1034–1050.
Spanioli P, Bommann C, Hauptmann G, Gerster T. 1996. Class III POU genes of zebrafish are predominantly expressed in the central nervous system. *Nucleic Acids Res* **24:** 4874–4881.
Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci* **102:** 15545–15550.
Thisse B, Thisse C. 2004. Fast release clones: A high throughput expression analysis. In *ZFIN Direct Data Submission* (http://zfin.org).
Thisse B, Heyer V, Lux A, Alunni V, Degrave A, Seiliez I, Kirchner J, Parkhill JP, Thisse C. 2004. Spatial and temporal expression of the zebrafish genome

by large-scale in situ hybridization screening. *Methods Cell Biol* **77:** 505–519.

Thompson JA, Ziman M. 2011. Pax genes during neural development and their potential role in neuroregeneration. *Prog Neurobiol* **95:** 334–351.

Vanacker JM, Pettersson K, Gustafsson JA, Laudet V. 1999. Transcriptional targets shared by estrogen receptor-related receptors (ERRs) and estrogen receptor (ER) α, but not by ERβ. *EMBO J* **18:** 4270–4279.

Verrijzer CP, Van der Vliet PC. 1993. POU domain transcription factors. *Biochim Biophys Acta* **1173:** 1–21.

Visel A, Minovitsky S, Dubchak I, Pennacchio LA. 2007. VISTA Enhancer Browser—a database of tissue-specific human enhancers. *Nucleic Acids Res* **35:** D88–D92.

Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, et al. 2009. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457:** 854–858.

Wang W, Zhong J, Wang YQ. 2010. Comparative genomic analysis reveals the evolutionary conservation of Pax gene family. *Genes Genet Syst* **85:** 193–206.

Waskiewicz AJ, Rikhof HA, Hernandez RE, Moens CB. 2001. Zebrafish Meis functions to stabilize Pbx proteins and regulate hindbrain patterning. *Development* **128:** 4139–4151.

Westerfield M. 2000. *The zebrafish book. A guide for the laboratory use of zebrafish* (Danio rerio), 4th ed. University of Oregon Press, Eugene, OR.

Received February 28, 2012; accepted in revised form June 21, 2012.

Curriculum Vitae

Xylena Reed
1935 Fleet St Apt. 2
Baltimore, MD 21231
xreed1@jhmi.edu
509-847-3650

Education:

Ph.D. Human Genetics. Johns Hopkins University School of Medicine, Baltimore, MD.
- January, 2015 (expected)
- Thesis: "Towards the genesis of neuronal regulatory catalogs and their vocabularies"

B.S. Biochemistry. University of Washington, Seattle, WA.
- June 2008
- GPA: 3.59/4
- Thesis: "Functional Studies of FRG1"

Research Experience:

2009-present. Johns Hopkins University School of Medicine, Baltimore, MD.
- Thesis advisor: *Andrew S. McCallion*, Ph.D.
- The focus of my research has been to develop and apply strategies for the discovery and characterization of enhancer sequences that drive expression in vertebrate neuronal populations, and ultimately to the specific analysis of dopaminergic neurons.
- Identified enhancers that drive expression in the central nervous system flanking *LMX1A*, *LMX1B* and *Phox2b* by using sequence conservation and reporter assays in zebrafish.
- Analyzed the biological significance of a classifier trained on 211 experimentally proven hindbrain enhancers finding that 88% of predicted enhancers tested showed reporter expression in stable lines of transgenic zebrafish. I further broke down a subset of these enhancers to reveal their transcription factor binding site clusters to show that they retained their enhancer function.
- Contributed significantly to the identification of 2489 putative melanocyte enhancers by completeing of ChIP-seq for H3K4me1 in cultured mouse melanocytes.
- I determined the feasibility of dissociation and FACS of the embryonic Th:GFP transgenic mouse midbrain for analysis by ChIP-seq. In order to use this population to generate a catalog of putative dopaminergic enhancers I have worked towards optimization of low cell number (<1M) ChIP-seq for histone modifications in our lab.

2008-2009. National Institute of Dental and Craniofacial Research, Bethesda, MD.
- Advisors: *Matthew P. Hoffman*, B.D.S., Ph.D. and *Sarah M. Knox*, Ph.D.
- I investigated the role of the parasympathetic nervous system in branching organ development using ex vivo culture of the embryonic mouse submandibular gland as a model system.

- Examined the effect of EGFR and muscarinic receptor agonists and antagonists on the expression of markers of epithelial and mesenchymal cells using quantitative RT-PCR.
- Characterized gene expression patterns in the submandibular gland using immunofluorescence and confocal microscopy.

2006-2008. University of Washington, Department of Biochemistry, Seattle, WA.

- Advisor: *Brian K. Kennedy*, Ph.D.
- I examined the function of FRG1 in cultured mouse myoblast to determine the possibility of its involvement in the pathogenesis of FSHD.
- Characterized the subcellular localization of overexpressed FRG1 mutants in cultured mouse myoblasts.
- Investigated possible binding partners of FRG1 using co-immunoprecipitations for protein and RNA.

Peer-reviewed publications:

Burzynski GM*, **Reed X***, Maragh S, Matsui T, McCallion AS. Integration of genomic and functional approaches reveals enhancers at LMX1A and LMX1B. *Mol Genet Genomics*. 2013 Aug. (*Co-first authors).

Gorkin DU, Lee D, **Reed X**, Fletez-Brant C, Bessling SL, Loftus SK, Beer MA, Pavan WJ, McCallion AS. Integration of ChIP-seq and machine learning reveals enhancers and a predictive regulatory sequence vocabulary in melanocytes. *Genome Research*. 2012 Nov.

Burzynski GM*, **Reed X***, Taher L*, Stine ZE, Matsui T, Ovcharenko I, McCallion AS. Systematic elucidation and in vivo validation of sequences enriched in hindbrain transcriptional control. *Genome Research*. 2012 Nov. (*Co-first authors).

Prasad MK, **Reed X**, Gorkin DU, Cronin JC, McAdow AR, Chain K, Hodonsky CJ, Jones EA, Svaren J, Antonellis A, Johnson SL, Loftus SK, Pavan WJ, McCallion AS. SOX10 directly modulates ERBB3 transcription via an intronic neural crest enhancer. *BMC Dev Biol*. 2011 Jun.

Chen SC, Frett E, Marx J, Bosnakovski D, **Reed X**, Kyba M, Kennedy BK. Decreased proliferation kinetics of mouse myoblasts overexpressing FRG1. *PLoS One*. 2011 May.

Knox SM, Lombaert IM, **Reed X**, Vitale-Cross L, Gutkind JS, Hoffman MP. Parasympathetic innervation maintains epithelial progenitor cells during salivary organogenesis. *Science*. 2010 Sep.

Presentations:

Reed X, Taher L, Burzynski GM, Fletez-Brant C, Lee D, Gary DS, Beer MA, Ovcharenko I, McCallion AS. Integrating functional and computational genomics to develop neuronal regulatory vocabularies. American Society for Human Genetics National Annual Meeting. Moscone Center, November 2012.

- Poster

Reed X, Knox SM, Hoffman MP. Identification of Genes Regulated by the Parasympathetic Ganglion during Submandibular Gland Branching Morphogenesis. Salivary Glands & Exocrine Secretion Gordon Research Conference. Hotel Galvez, February 2009.

- Poster

Reed X, Marx JG, Kennedy BK. Functional Studies of FRG1. Mary Gates Undergraduate Research Symposium. University of Washington, May 2007.

- Undergraduate research talk

Additional Meetings and Symposia Attended:

2013 - Epigenetics in Development. The Center of Excellence in Chromosome Biology, National Cancer Institute. Bethesda, MD.
2012 - Epigenomics of Common Diseases. Wellcome Trust Conference. Johns Hopkins University, Baltimore, MD.
2011 - Chromatin Structure and Function. The Center of Excellence in Chromosome Biology, National Cancer Institute. Bethesda, MD.

Scientific Courses Attended:

2014 - Molecular Neurology and Therapeutics. Wellcome Trust Advanced Course. Wellcome Trust Genome Campus, Hinxton, Cambridge, UK.
2010 - Medical and Experimental Mammalian Genetics. The Jackson Laboratory. Bar Harbor, ME.

Mentoring and Teaching Experience:

2014 - Mentored first year graduate student

- Instructed first year Ph.D. student on conceptual and technical lab skills for analysis of neuronal populations in mouse and zebrafish.

2012 - Teaching assistant for Genetics and Medicine: History of Ideas

- Advised first year Ph.D. students on presentations and discussed significant journal articles regarding the evolving concept of a gene.

Honors and Awards

2014 - Bursary award for the Wellcome Trust Advanced Course in Molecular Neurology and Therapeutics, Wellcome Trust
2010-2013 - Faculty of 1000 Associate Faculty Member
2009-2010 - Post Baccalaureate Intramural Research Training Award, National Institutes of Health
2008 - Biochemistry Departmental Distinction, University of Washington
2007-2008 - Mary Gates Research Scholarship, University of Washington
2005-2008 - University of Washington Dean's List